OPINION EXTRACTION AND SUMMARIZATION FROM TEXT DOCUMENTS IN BENGALI

THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY (ENGINEERING)

OF

JADAVPUR UNIVERSITY

AMITAVA DAS

Department of Computer Science & Engineering

Jadavpur University

Kolkata 700032

December 2011

JADAVPUR UNIVERSITY

Kolkata – 700032, India

INDEX NO. 101/08/E

1. Title of Thesis:

Opinion Extraction and Summarization from Text Documents in Bengali

2. Name, Designation & Institution of the Supervisor/s:

Prof. (Dr.) Sivaji Bandyopadhyay
Department of Computer Science & Engineering,
Jadavpur University, Kolkata-700 032, India

3. List of Publications (33):

A. Book Chapters (2):

- **Amitava Das** and Sivaji Bandyopadhyay. 2012(a). ***Sentiment…Human Intelligence***. *In Online Collective Action Dynamics of the Crowd in Social Media as a Book Chapter. Springer Lecture Notes in Social Networks*. (**Accepted**)

- Asif Ekbal, Sivaji Bandyopadhyay and **Amitava Das.** 2007. ***Three Different Models for Named Entity Recognition in Bengali.*** *In the Proceedings of the International Workshop on Advances in Pattern Recognition (IWAPR-2007), Springer (Advances in Pattern Recognition Series)*, Pages 161-170, UK.

  http://books.google.com/books?id=KflvrMCXGYQC&lpg=PA161&ots=4trkG_RurZ&dq=Amitava%20Das%2Bnlp&pg=PA161#v=onepage&q=Amitava%20Das+nlp&f=false

B. Journal Publications (3):

- **Amitava Das** and Sivaji Bandyopadhyay. 2010(a). ***Phrase-level Polarity Identification for Bengali.*** *In International Journal of Computational Linguistics and Applications (IJCLA)***,** *Vol. 1, No. 1-2, ISSN 0976-0962*, Pages 169-182, Jan-Dec 2010.
  http://www.ijcsit.com/docs/Volume%202/vol2issue1/ijcsit2011020107.pdf

- **Amitava Das** and Sivaji Bandyopadhyay. 2010(b). ***Syntactic Sentence Fusion Techniques for Bengali***. *In International Journal of Computer Science and Information Technologies (IJCSIT)*, ISSN: 0975-9646. Vol. 2 (1), 2011, Pages 494-503.
  http://www.ijcsit.com/docs/Volume%202/vol2issue1/ijcsit2011020107.pdf

- **Amitava Das** and Sivaji Bandyopadhyay. 2012(b). ***Opinion is the Medium between knowledge and Ignorance***. *In International Journal of Computer Science and Information Technology*. (**Accepted**).

C. Conference/Workshop Publications (Chapter Wise) (28):

SENTIMENT LEXICON ACQUISITION (CHAPTER 1) (6)

1. **Amitava Das** and Sivaji Bandyopadhyay. 2011. ***Dr Sentiment Knows Everything!*** *In the Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011 Demo Session)*, Pages 50-55, Portland, Oregon, USA.
   http://www.aclweb.org/anthology/P11-4009

2. **Amitava Das.** 2011. ***PsychoSentiWordNet***. *In the Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011 Student Session)*, Pages 52-57 Portland, Oregon, USA.
   http://aclweb.org/anthology/P/P11/P11-3010.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2010(c). ***Dr Sentiment Creates SentiWordNet(s) for Indian Languages Involving Internet Population***. *In the Proceeding of IndoWordNet Workshop, ICON 2010*, Kaharagpur, India.
   http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/14_iwn_SentiWordNet%28s%29%20for%20Indian%20Languages.pdf

4. **Amitava Das** and Sivaji Bandyopadhyay. 2010(d). ***Towards The Global SentiWordNet***. *In the Proceeding of the Workshop on Model and Measurement of Meaning (M3), PACLIC 24th*, Pages 799-808, Sendai, Japan.
   http://www.amitavadas.com/Pub/GSN.pdf

5. **Amitava Das** and Sivaji Bandyopadhyay. 2010(e). ***SentiWordNet for Indian Languages***. *In the Proceeding of the 8th Workshop on Asian Language Resources (ALR 8), COLING 2010*, Pages 56-63. Beijing, China.
   http://aclweb.org/anthology/W/W10/W10-3208.pdf
   http://59.108.48.12/proceedings/coling/coling2010/ALR/pdf/ALR08.pdf

   CITED BY 1

   I. Quang-Thuy Ha, Tien-Thanh Vu, Huyen-Trang Pham and Cong-To Luu. 2011. ***An Upgrading Feature-Based Opinion Mining Model on Vietnamese Product Reviews***. *In the Proceeding of the 7th International Conference on Active Media Technology (AMT 11)*, Pages 173-185, Berlin, Heidelberg.
   http://dl.acm.org/citation.cfm?id=2033921

6. **Amitava Das** and Sivaji Bandyopadhyay. 2010(f). ***SentiWordNet for Bangla***. *In the Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary (KSE4)*, Mysore, India.
http://www.ldcil.org/up/conferences/Dictionary/Dict-Event_Schedule.pdf
http://www.amitavadas.com/Pub/SentiwordNet%20%28Bengali%29.pdf

SENTIMENT / SUBJECTIVITY DETECTION (CHAPTER 2) (4)

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010(g). ***Subjectivity Detection using Genetic Algorithm***. *In the Proceeding of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA10), ECAI 2010*, Pages 14-21, Lisbon, Portugal.
http://www.amitavadas.com/Pub/GA.pdf

2. **Amitava Das** and Sivaji Bandyopadhyay. 2009(a). ***Subjectivity Detection in English and Bengali: A CRF-based Approach***. *In the Proceeding of the International Conference on Natural Language Processing (ICON 2009)*, Pages 358-363, Hyderabad, India.
www.amitavadas.com/Pub/ICON_Final_Amitava.pdf

CITED BY 5

I. Smruthi Mukund and Rohini K Srihari. 2010. ***A Vector Space Model for Subjectivity Classification in Urdu Aided By Co-Training***. *In the Proceeding of the 23rd International Conference on Computational Linguistics (COLING '10)*, Pages 860—868, Beijing, China.
http://aclweb.org/anthology/C/C10/C10-2099.pdf

II. Matthew Kusner. 2011. ***Emotional Feedback Generation for Physical Therapy***. *Honors Projects, Paper 23*, Macalester College, Minnesota, USA.
http://digitalcommons.macalester.edu/mathcs_honors/23/

III. Malik Atalla, Christian Scheel, Ernesto William De Luca and Sahin Albayrak. 2011. ***Investigating the Applicability of current Machine-Learning based Subjectivity Detection Algorithms on German Texts***. *In the Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, Pages 17–24, Hissar, Bulgaria.
http://www.aclweb.org/anthology-new/W/W11/W11-3903.pdf

IV. K M Azharul Hasan, Esmot Ara, Farhana Hoque, Jenifar Yasmin. 2010. ***A Multidimensional Partitioning Scheme for Developing English to Bangla Dictionary***. *In the Proceedings of the 11th International Conference on Computer and Information Technology (ICCIT 2010)*, Pages 92-96, Dhaka, Bangladesh.
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5723835

V. Aditya Joshi, Balamurali A R and Pushpak Bhattacharyya. 2010. ***A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study***. *In the Proceedings of the 8th ICON 2010*, Kharagpur, India.
http://www.cse.iitb.ac.in/~balamurali/papers/ICON%20229.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2009(b). ***Theme Detection an Exploration of Opinion Subjectivity.*** *In the Proceeding of the Affective Computing & Intelligent Interaction (ACII2009)*, Pages 1-6, Amsterdam, Netherlands.
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5349599

CITED BY 2

I. Ankur Rana, Vishal Goyal and Vimal K. Soni. 2011. ***Renaissance of Opinion Mining***. *In Information Systems for Indian Languages, Communications in Computer and Information Science, Volume 139, Part 1,* Pages 60-67.
www.springerlink.com/index/L570K513R2743L76.pdf

II. Pengcheng Han, Junping Du and Liping Chen. 2010. ***Web Opinion Mining Based On Sentiment Phrase Classification Vector***. *In the Proceeding of the 2nd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC 2010)*, Pages 308-312, Beijing, China.
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5657968

4. **Amitava Das** and Sivaji Bandyopadhyay. 2009(c). ***Extracting Opinion Statements from Bengali Text Documents through Theme Detection.*** *In the Proceeding of the 17th International Conference on Computing (CIC-09), GEOS 2009*, Mexico City, Mexico.
www.amitavadas.com/Pub/CIC_Poster_2008_JU.pdf

SENTIMENT POLARITY CLASSIFICATION (CHAPTER 3) (2)

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010(h). ***Opinion-Polarity Identification in Bengali***. *In the Proceeding of the 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL 2010)*, *KESE 2010*, Pages 41-44, Redwood City, California, USA.
www.amitavadas.com/Pub/ICCPOL_2010.pdf

2. **Amitava Das** and Sivaji Bandyopadhyay. 2012(c). ***Sentimantics: The Conceptual Spaces for Human Cognition and Sentiment***. *In the Proceeding of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, Delhi, India. (**Communicated**).

SENTIMENT STRUCTURIZATION (CHAPTER 4) (1)

1. **Amitava Das**, Aniruddha Ghosh and Sivaji Bandyopadhyay. 2010(i). ***Semantic Role Labeling for Bengali Noun using 5Ws: Who, What, When, Where and Why***. *In the Proceeding of the International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLPKE2010)*, Pages 1-8, Beijing, China, 2010.
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5587772

SENTIMENT SUMMARIZATION-VISUALIZATION-TRACKING (CHAPTER 5) (3)

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010(j). ***Opinion Summarization in Bengali: A Theme Network Model***. *In the Proceeding of the 2nd IEEE International Conference on Social Computing (SocialCom-2010)*, Pages 675-682, Minneapolis, USA.
   http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5591520

2. **Amitava Das** and Sivaji Bandyopadhyay. 2010(k). ***Topic-Based Bengali Opinion Summarization***. *In the Proceeding of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Pages 232-240, Beijing, China.
   http://aclweb.org/anthology/C/C10/C10-2027.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2012(d). ***The 5W Structure for Sentiment Summarization-Visualization-Tracking***. *In the Proceeding of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, Delhi, India. (**Accepted**).

APPENDIX: STEMMER AND DEPENDENCY PARSER (3)

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010(l). ***Morphological Stemming Cluster Identification for Bangla*.** *In Knowledge Sharing Event-1: Task 3: Morphological Analyzers and Generators*, January 24-25, 2010, Mysore, India.
   http://www.amitavadas.com/Pub/Morph_Bengali.pdf

2. Aniruddha Ghosh, **Amitava Das**, Pinaki Bhaskar and Sivaji Bandyopadhyay. 2010(m). ***Bengali Parsing System at ICON NLP Tool Contest 2010***. *In the NLP Tool Contest*: *Dependency Parsing at International Conference on Natural Language Processing (ICON 2010),* Pages 20-24, Kharagpur, India.
   http://researchweb.iiit.ac.in/~prashanth/papers/husain-mannem-ambati-gadde-icon10.pdf
   http://www.sivajibandyopadhyay.com/pinaki/papers/JU%20parsing%20system-2010.pdf

3. Aniruddha Ghosh, **Amitava Das**, Pinaki Bhaskar and Sivaji Bandyopadhyay. 2009(d). ***Dependency Parser for Bengali: the JU System at ICON 2009***. *In the NLP Tool Contest*: *Dependency Parsing at International Conference on Natural Language Processing (ICON 2009)*, Pages 7-11, Hyderabad, India.
   http://ltrc.iiit.ac.in/nlptools2009/CR/all-papers-toolscontest.pdf
   http://ltrc.iiit.ac.in/nlptools2009/CR/Parser_Camera_Ready_JU.pdf

MISCELLANEOUS (9)

1. **Amitava Das**. 2010(n). ***Can We Mimic Human Pragmatics Knowledge into Computational Lexicon?*** *In the International Conference on Natural Language Processing (ICON 2010)*, December 8-11, Kharagpur, India.
   Best Student Paper Award
   http://ltrc.iiit.ac.in/icon_archives/ICON2010/10Dec2010/Amitav-Camera-Ready-Copy-Student-Paper-Comp.pdf

2. Aniruddha Ghosh, **Amitava Das** and Sivaji Bandyopadhyay. 2010(o). ***Clause Identification and Classification in Bengali.*** *In the Proceeding of the Workshop on South and Southeast Asian Natural Language Processing (SSANLP) - Morphology and Segmentation, COLING 2010*, Pages 17-25, Beijing, China.
   http://aclweb.org/anthology/W/W10/W10-3603.pdf

   CITED BY 1

   I. Daraksha Parveen, Ratna Sanyal and Afreen Ansari. 2011. ***Clause Boundary Identification using Classifier and Clause Markers in Urdu Language***. *In Research Journal on Computer Science and Computer Engineering with Applications (POLIBITS)*, Pages 61-65.
   http://www.cidetec.ipn.mx/polibits/2011_43/43_08.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2010(p). ***SemanticNet-Perception of Human Pragmatics***. *In the Proceeding of the 2nd Workshop on Cognitive Aspects of the Lexicon: Enhancing the Structure and Lookup Mechanisms of Electronic Dictionaries (COGALEX-II), COLING 2010*, Pages 2-11, Beijing, China.
   http://aclweb.org/anthology/W/W10/W10-3400.pdf

4. Asif Ekbal, Rejwanul Haque, **Amitava Das**, Venkateswarlu Poka and Sivaji Bandyopadhyay. 2008(a). ***Language Independent Named Entity Recognition in Indian Languages.*** *In the Proceedings of the Workshop on Named Entity Recognition for South and South East Asian Languages, IJCNLP-08*, Pages 33-40, Hyderabad, India.
   www.mt-archive.info/IJCNLP-2008-Ekbal.pdf

   CITED BY 8

   I. Junwu Du, Zhimin Zhang, Jun Yan, Yan Cui and Zheng Chen. 2010. ***Using Search Session Context for Named Entity Recognition in Query***. *In the Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pages 765-766, New York, USA.
   http://59.108.48.12/proceedings/SIGIR/SIGIR2010/docs/p765.pdf

   II. Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar and Pabitra Mitra. 2008. ***A Hybrid Approach for Named Entity Recognition in Indian Languages***. *In the IJCNLP-08*, Pages 17–24, Hyderabad, India.
   http://ltrc.iiit.ac.in/ner-ssea-08/drafts/5.pdf

III. Pramod Kumar Gupta, Sunita Arora. 2009. ***An Approach for Named Entity Recognition System for Hindi: An Experimental Study***. *In the Proceedings of the Second Annual Seminar of CDAC Noida Technologies (ASCNT-09)*, Pages 103-108, Noida, India.
http://220.156.188.21/CDAC/ASCNT_2009/ASCNT%202009/Paper/Language%20computing/Abstract10.pdf

IV. Kashif Riaz. 2010. ***Rule-Based Named Entity Recognition in Urdu***. 2010. In the Proceedings of the Named Entities Workshop (NEWS 2010), ACL 2010, Pages 126-135, Uppsala, Sweden.
http://aclweb.org/anthology/W/W10/W10-2419.pdf

V. Rajesh Sharma and Vishal Goyal. 2011. ***Name Entity Recognition Systems for Hindi Using CRF Approach***. *In Information Systems for Indian Languages Communications in Computer and Information Science, Volume 139, Part 1,* Pages 31-35.
http://www.springerlink.com/content/l0847866p60k42g8/

VI. Padmaja Sharma, Utpal Sharma and Jugal Kalita. 2011. ***Named Entity Recognition: A Survey for the Indian Languages***. *In the Problems of Parsing in Indian Languages*, Pages 35-40.
http://languageinindia.com/may2011/kommaluricomplete.pdf#page=41

VII. Bindu M.S and Sumam Mary Idicula. 2011. ***Named Entity Recognizer employing Multiclass Support Vector Machines for the Development of Question Answering Systems***. *In International Journal of Computer Applications 25(10):40-46*, New York, USA.

http://www.ijcaonline.org/archives/volume25/number10/3146-4343

VIII. Ram, R.V.S., Akilandeswari, A., Devi, S.L. 2010. ***Linguistic Features for Named Entity Recognition Using CRFs***. *In the Proceeding of the International Conference on Asian Language Processing (IALP 2010)*, Pages 158-161, Harbin, China.
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5681603

5. **Amitava Das**, Tanik Saikh, Tapabrata Mondal and Sivaji Bandyopadhyay. 2010n. ***JU_CSE_GREC10: Named Entity Generation at GREC 2010.*** *In the Proceeding of the Generation Challenges on Named Entity Generation (GREC-NEG 2010), INLG 2010*, Pages 235-236, Dublin, Ireland.
www.aclweb.org/anthology/W10-4229

6. **Amitava Das**, Tanik Saikh, Tapabrata Mondal, Asif Ekbal and Sivaji Bandyopadhyay. 2010(q). ***English to Indian Languages Machine Transliteration System at NEWS 2010***. *In the Proceeding of the NEWS 2010, ACL 2010*, Pages 71-75, Uppsala, Sweden.
Stood 1st for English-Bengali Transliteration
www.mt-archive.info/NEWS-2010-Das.pdf

CITED BY 1

I. Jiampojamarn, Sittichai. 2011. ***Grapheme-to-phoneme conversion and its application to transliteration***. PhD Thesis. University of Alberta. https://www.cs.ualberta.ca/news-events/event-calendar/2010/grapheme-phoneme-conversion-and-its-application-transliteration

7. Pinaki Bhaskar, **Amitava Das**, Partha Pakray and Sivaji Bandyopadhyay. 2010(r). ***Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010***. *In the Forum for Information Retrieval Evaluation (FIRE-2010)*, Gandhinagar, India. http://www.isical.ac.in/~fire/paper_2010/Pinaki-Fire-2010_SB_PB_AD_PP.pdf

8. **Amitava Das**, Asif Ekbal, Tapabrata Mondal and Sivaji Bandyopadhyay. 2009(e). ***English to Hindi Machine Transliteration at NEWS 2009.*** *In the Proceeding of the Named Entity Workshop (NEWS 2009),* ACL-IJCNLP 2009, Pages 80-83, Singapore.
Stood 3rd for English-Hindi Transliteration
http://aclweb.org/anthology-new/W/W09/W09-3517.pdf
www.mt-archive.info/**NEWS-2009**-Das.pdf

CITED BY 1

I. Jiampojamarn, Sittichai. 2011. ***Grapheme-to-phoneme conversion and its application to transliteration***. *PhD Thesis*. University of Alberta. https://www.cs.ualberta.ca/news-events/event-calendar/2010/grapheme-phoneme-conversion-and-its-application-transliteration.

9. Sivaji Bandyopadhyay, **Amitava Das** and Pinaki Bhaskar. 2008(b). ***English Bengali Ad-hoc Monolingual Information Retrieval Task Result at FIRE 2008.*** *In the Forum for Information Retrieval Evaluation (FIRE-2008)*, Kolkata, India. www.isical.ac.in/~fire/paper/Bandhyopadhyay-JU-fire2008.pdf

4. List of Patents:

None

5. List of Presentations in National / International (11):

- **3RD INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (IJCNLP-08), HYDERABAD, INDIA**
  - ✓ *Language Independent Named Entity Recognition in Indian Languages.* ***(Oral)***
- **FORUM FOR INFORMATION RETRIEVAL EVALUATION (FIRE-08), KOLKATA, INDIA**
  - ✓ *English Bengali Ad-hoc Monolingual Information Retrieval.* ***(Oral)***
- **ACL_IJCNLP-2009, SINGAPORE**
  - ✓ *English to Hindi Machine Transliteration at NEWS 2009.* ***(Poster)***

- **7TH INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING (ICON-2009), HYDERABAD, INDIA**
  - ✓ *Detection in English and Bengali: A CRF-based Approac* ***.(Poster)***
- **THE 23RD INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING 2010), BEIJING, CHINA**
  - ✓ *SentiWordNet for Indian Languages*. ***(Oral)***
  - ✓ *SemanticNet-Perception of Human Pragmatics*. ***(Oral)***
  - ✓ *Clause Identification and Classification in Bengali*. ***(Oral)***
  - ✓ *Topic-Based Bengali Opinion Summarization*. ***(Poster)***
- **THE 6TH IEEE INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND KNOWLEDGE ENGINEERING (IEEE NLP-KE'10), BEIJING, CHINA**
  - ✓ Semantic Role Labeling for Bengali Noun using 5Ws**.** ***(Oral)***
- **THE 8TH INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING (ICON 2010), IIT KHARAGPUR, INDIA**
  - ✓ Can We Mimic Human Pragmatics Knowledge into Computational Lexicon? **(Oral)**
- **THE 3RD INDOWORDNET WORKSHOP, KHARAGPUR, INDIA**
  - ✓ Dr Sentiment Creates SentiWordNet(s) for Indian Languages Involving Internet Population**. (Oral)**

Certificate From The Supervisor

This is to certify that the thesis entitled "**Opinion Extraction and Summarization from Text Documents in Bengali**" submitted by **Mr. Amitava Das**, who got his name registered on 8th September, 2008, for the award of Ph.D. (Engineering) degree of Jadavpur University is absolutely based upon his work under the supervision of Prof. (Dr.) Sivaji Bandyopadhyay and that neither his thesis nor any part of the thesis has been submitted for any degree or any other academic award anywhere before.

Signature of the Supervisor

Prof. (Dr.) Sivaji Bandyopadhyay

Department of Computer Science & Engineering,

Jadavpur University,

Kolkata-700032, India

Declaration

I declare that the work described in this thesis is entirely my own. No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or institute. Any help or source information, which has been availed in the thesis, has been duly acknowledged.

Signature

Amitava Das

Department of Computer Science & Engineering,

Jadavpur University,

Kolkata-700032, India

To My ***Parents....***

ACKNOWLEDGEMENTS

Throughout the period of becoming a doctor of philosophy, this is the only occasion when one could be completely free with his philosophy and thoughts. Please excuse my language errors; in another way these errors may help you understand the unabridged 'Amitava'. But, if you feel you had enough of the errors already or you easily get bored with philosophical conundrums then this section is not for you in the first place, please go ahead and look forward to the rest of the 200 pages in this dissertation.

Who am I?

After my graduation when I refused to go for the conventional dream jobs that were considered to be best options to become a quick success, I decided to step into the uncharted world of research. The first person to stand behind me is my father. He said "*Look Son, I was compelled to leave my academics due to the fact that I could not pay my school fees on time, but come what may, I will be there for you to help you all the time and by all means possible. Do what you love to do, go ahead and pursue your career in research*." He kept his word verbatim. My mother is solely responsible for my attitude development. She teaches me to be দুঃখেষুঅনুদ্বিগ্নমনা সুখেসুবিগতস্পৃহা ('one who is not overwhelmed by happiness neither gets washed away by sorrow' in other words 'don't celebrate your success too much, failure is all that might be waiting for you next step'). I really do not know how much I become able to follow her teachings to become an indifferent person to this stature. This attitude helps me a lot to sustain in the professional field and pursue further. In detail I never get overwhelmed by my success because I know there is a failure at very near corner waiting for me and on the contrary from a failure I acquire a deep breath to run even faster to reach the unreachable horizon of success. My sister always encourages me in a discouraging manner. She usually tells me not to study because it inflicts an indirect pressure on her to study too. But actually she takes a lot of pride in discussing about me to her friends. Moreover she is the only person who never thinks twice to support me because for her "*Dadabhai is always right.*".

This is for my better half Suranjana. She is more interested in my paintings rather than my research. She taught me a great lesson of life and for that I am thankful to her and I respect her as a person. After receiving the best paper award at ICON in 2010 I called her to share my joy but eventually she was ice cool and with no joyful reaction. After a few hours I again called her and asked "*Are you not happy?*". She replied "Yes *I am, celebrate your success with your family, friends, colleagues, superiors and me also but do remember failures will come into life and nobody will be with you and then I will be there for you and for always.*" This is the best sentimental reflection for me, the sentiment analysis researcher.

This is for my beloved friends. Though I do not believe in luck but whenever I look at my friends I feel lucky. First of all, it is about Pinaki, who is my friend since my childhood days. From those days till today we are together and presently doing research in the same laboratory. The friendship is a great achievement for both of us. Another important friend is Krishnendu. Though he is a friend but I respect him as a person very much. In 2006 when I told him that I will do research he asked me "*For more money or more vibrant job? Or for true knowledge?*". I replied "*For the second one.*". He said "*Go for it.*". There are so many other friends and I learned many things from them: Atanu Da: Simplicity, Suman: Hard Target, Arghya: Enthusiasm, Debdeep: Western Culture, Rationality and English. Thanks Mrinmoy for your Credit card support many times. If I miss some name please do mind and tell me.

The world helps me to become a researcher.

First of all, Asif Da, took me to my laboratory at Jadavpur University from Murshidabad College of Engineering and Technology (MCET) in 2006. Moreover he is the first person who pampered my research attitude for which I became what I am today. Thanks to Asif Da.

I would like express my profound gratitude for Prof. Sivaji Bandyopadhyay, who taught me "*How to do research*". Five years back when I joined here I was only a programmer and had no knowledge of paper writing, research directions and etc. With time he taught me everything that I need to learn to complete my PhD, which could not be possible without his caring and grooming.

Dipankar, thanks for giving me a nice competition, which instigated the "*Struggle for Existence*" situation, it helped me a lot to complete my PhD at right time. Thanks for being my co-passenger in every foreign trip. Though you became my co-passenger not by choice but rather by chance but still the overall experience with you is good. Let us see when our destinies bring us face to face again.

This is for Prof. Pushpak Bhattacharya, IIT Bombay. I met him first time in 2007 at Jadavpur University. He delivered a nice scientific lecture along with a touch of philosophy: "*There are two basic types of PhDs, one who runs for paper and the second, who do research first and then publish it.*". I followed the second path and probably I succeeded. Anyone can see my publication rates 2008: 2 papers, 2009: 5 papers, 2010: 18 papers. It does not mean that I did the whole research in 2010. Rather, I did it earlier but published it in 2010. Another important aspect of his personality is his simplicity. Whenever he visits JU he always asks everybody about him or his family and even Rashed Da, who is merely a peon in our laboratory. It shows his kind heartedness and humanity.

The Five years of life here in Jadavpur University

As I talked about Rashed Da then I should thank to him. But not for giving me coffee in my bed at morning or not for the other mundane household jobs, it is his duty and he is paid for it! But I

do believe nobody knows Rashed Da as a poet! His sense of language is very poor due to his education standard but sometimes the contents awake hearts. Once he wrote a poem to satire or ridicule us, the mess members. I can't quote his line here directly but the narrative meaning was "*though you are earning a lot but still you are complaining! Look at me I have nothing and there is no hope of having great in future ahead, but still I am happy!*". It reminds me the curious case of the beggar in the story of "রাজা অসুখ" (*King's Disease)* by Sukumar Roy.

Finally, I would like to thank all my laboratory fellows for personal or professional helps. Especially, Anup Da for lending me money twice, when I badly needed it. Thanks everybody, Sudip Da, ParthaDa, Santanu Da, Tapabrata, Tanik and others.

Please do not mind by orders of names. All errors are - of course - mine.

(AMITAVA DAS)

ABSTRACT

Sentiment analysis or opinion mining refers to the application of natural language processing, computational linguistics and text analytics to identify and extract sentimental/opinionated/emotional information from text. Actually sentiment analysis, opinion mining and emotion analysis refers to the same task. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed sentiment in a document, a sentence or an entity feature/aspect is positive (*happy*), negative (*sad*) or neutral (*memorable*). The feature/aspect level analysis of sentiment demands proper structures for more precise sentiment extraction. Sentiment/opinion aggregation is a necessary requirement at the end users' perspective. Therefore summarization is necessary to present an at-a-glance presentation of the main points made in a single sentiment/opinion or how sentiment/opinion changes from time to time over multiple documents.

Sentiment Analysis (SA) from natural language text is a multifaceted and multidisciplinary problem simultaneously. SA defines an overall problem, which address multiple aspects of sub-problems. Human sentiment knowledge grows with its age and daily cognitive interactions. Therefore an intelligent human should need some prior knowledge to act properly. Sentiment knowledge acquisition is generally wrapped into computational lexicon, technically called **Sentiment Lexicon**. Similar to classical pattern recognition problems, SA is also classified into identification and classification genre called **subjectivity detection** and **polarity classification** that involve sentiment detection and sentiment classification. Proper **structurization** is required to proceed for any further granular analysis. Structurization involves identification of sentiment holder, sentiment topic and so on. The philosophical notion of science is always: "Necessity is the mother of all invention". Therefore the information processing need drives us to develop such systems that should meet the user satisfaction level. Therefore textual or visual **summarization** or **tracking** of sentiment is the striking need for the end user. The overall experiments described in the present thesis mainly deals with **English** or **Bengali** languages or both.

The present thesis is distributed in five chapters. The first chapter describes the Sentiment knowledge acquisition process in terms of **Sentiment Lexicon** while the second and third chapter describes the **subjectivity detection** and **polarity classification** problem respectively. To produce a formidable output for the end user proper structurization is required. The **structurization** has been described in the chapter four and finally the **Summarization-Visualization-Tracking** methodologies are described in the chapter five. In the conclusion section, the key contributions of the present work in the Sentiment Analysis research have been self explained. In the *Appendix* section the details of some developed natural language processing (NLP) tools like *Dependency Parser* and *Stemmer* for Bengali are reported. The following paragraphs give a very brief description about the chapters in the present thesis.

Sentiment knowledge acquisition in terms of sentiment lexicon is the vital pre-requisite of any sentiment analysis system. Previous studies have proposed to attach **prior polarity** to each sentiment lexicon level. A number of research endeavors could be found in the literature for creation of sentiment

lexicon in several languages and domains. These techniques can be broadly categorized into two genres, one follows the classical manual annotation techniques and the other includes various automatic techniques. Both types of techniques have few limitations. Manual annotation techniques are undoubtedly trustable but it generally takes time for development. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain. Manual annotation techniques require a large number of annotators to balance the sentimentality of individual annotators in order to reach agreement. But qualified human annotators are quite unavailable and also costly. Both the processes have been attempted to develop ***sentiment lexicon*** for multiple languages. The automatic processes used in the present work are ***bilingual dictionary based approach, WordNet based synonym and antonym expansion, orthographic antonym generation*** and ***corpus based approach***. These automatic techniques have been successfully applied for three Indian languages: **Bengali, Hindi** and **Telugu.** As there is high scarcity of human annotators, it has been decided to involve the **Internet Population** for creating more credible sentiment lexicons. An online game called **Dr Sentiment** has been developed which is a template based interactive online game that collects players' sentiment by asking a set of simple template based questions. The lexicons tagged by this system are credible as it is tagged by human beings. It is not a static sentiment lexicon set as the prior polarity scores are updated regularly. **Global SentiWordNet** for **57 languages** has been developed by Dr Sentiment. Moreover the online game helps to collect several psychological information along with the sentiment knowledge and the resultant lexicon is termed as the **PsychoSentiWordNet**. The PsychoSentiWordNet holds variable prior polarity scores that may be fetched depending upon the regulating psychological aspects like location, age, gender, profession etc.

The term subjectivity simply refers to the identification of sentiments in a piece of text. More precisely, the term ***Subjectivity*** can be defined as the ***Topical Relevant Sentiment.*** The subjectivity is concerned with whether the expressed sentiment is related to the relevant topic or it fulfills the overall desired goal of a Sentiment Analysis system. The subjectivity experiments started with the **Rule-based** technique and continued with **Machine Learning** and **Hybrid** techniques. The Theme Detection technique has been developed to detect topical relevant sentiments. The themes relate to the sentimental topic of any document. But there may be more unrevealed clues based on human psychology or on complex relationships among the linguistic clues for sentiment / subjectivity detection which may not be extracted with present NLP/simple machine learning techniques. Thus, experiments have been carried out with **Genetic Algorithm** to adopt the biological evolutionary path of the human intelligence for machines. The accuracy of the system with Genetic-Based Machine Learning (GBML) based technique reaches **90.22%** (MPQA: news), **93.00%** (IMDB: movie review) and **87.65%** (news) **90.6%** (blog) for English and Bengali respectively

The polarity classification task involves sentiment/opinion classification into semantic classes such as *positive, negative or neutral* and/or other fine-grained emotion classes like *happy, sad, anger, disgust and surprise* The two step methodology, i.e., use of prior polarity lexicon followed by any NLP technique is the standard method for the polarity classification task as established by several previous research efforts. The SentiWordNet (Bengali) (discussed in the Chapter One) developed as part of the present work has been used as the prior polarity lexicon. The application of the NLP techniques started with the

Syntactic-Statistical classification technique. The Support Vector Machine (SVM) has been used with a number of features for the development of the syntactic polarity classifier. The polarity classification task mainly involves **syntactic analysis** like **Modifier-Modified** relationship or **Association ambiguity**. A **dependency parser** has been developed for Bengali language as there was no dependency parser available. The polarity classification performance of the Syntactic polarity classifier reaches 70.04%. Dealing with unknown/new words is a common challenge for NLP systems. It becomes more difficult for sentiment analysis because it is very hard to find out any contextual clue to predict the sentimental orientation on any unknown/new word. A prior polarity lexicon is attached with two probabilistic values, i.e., positivity and negativity scores and there is no clue in the SentiWordNet regarding ***which value to pick in what context?.*** The general trend is to pick the one with the highest score but that may vary depending on the context. For example, the word "***High***" (Positivity: 0.25, Negativity: 0.125 for "***High***" from the SentiWordNet) is attached with a positive (positivity value is higher than the negativity value) polarity in the sentiment lexicon but the polarity of the word may vary. Additional NLP techniques are required to disambiguate these types of words. There are 6619 lexicon entries in the English SentiWordNet where both the positivity and the negativity values are greater than zero. Similarly there are a total of 17927 lexical entries in the English SentiWordNet, whose positivity and negativity value difference is less than 0.2. These entries are ambiguous because there is no clue in the SentiWordNet regarding the positivity or negativity of such entries. The research attempts in the present work are mainly concerned about the ambiguous entries from the SentiWordNet. The basic hypothesis is that if some sort of contextual information can be added in the sentiment lexicon along with the prior polarity scores then the updated rich lexicon network will serve better than the existing one and it may lessen the requirements for further NLP techniques to disambiguate the contextual polarity. A new paradigm called ***Sentimantics*** has been introduced in the present work which uses distributed Semantic Lexical Models to hold the sentiment knowledge with contextual common sense. To the best of our knowledge such a paradigm has not been explored before.

The need of the end user is the driving force behind the sentiment analysis research. The outcomes of these research endeavors should lead to the development of a real time sentiment analysis system, which will successfully satisfy the need of the end users. Let us have a look at some real life needs of the end user. For example, a market surveyor from company A may identify the need to find out the changes in public opinion about their product X after release of product Y by another company B. The different aspects of product Y that the public consider better than product X are also points of interest. These aspects could typically be the durability of the product, power options, weight, color and many more other issues that depend on the particular product. In another scenario, a voter may be interested to study the change of public opinion about any leader or any public event before and after any election. In this case the aspect could be a social event, economic recession and may be other issues. The end users are not only looking for the binary (positive/negative) sentiment classification but they are more interested in aspectual sentiment analysis. Therefore, only sentiment detection and classification is not enough to satisfy the need of the end user. A sentiment analysis system should be capable enough to understand and extract out the aspectual sentiments present in a natural language text. Previous research efforts have already proposed various structures or components for sentiment extraction. Among the proposed sentiment structures the most widely used structures are **Holder**, **Topic** and other

domain dependant **Attributes**. But the real life users are not always interested about all the aspects at a time, rather they look for opinion/sentiment changes of any "Who" during "When" and depending upon "What" or "Where" and "Why". With this hypothesis, the 5W (**Who**, **What**, **When**, **Where** and **Why**) constituent extraction technique for sentiment/opinion structurization has been proposed. The proposed 5W structure is domain independent and more generic than the existing semantic constituent extraction structure.

Aggregation of information is the necessity from the end users' perspective but it is nearly impossible to develop consensus on the output format or how the data should be aggregated. An end user might want to have an at-a-glance presentation of the main points made in a single review or how opinion changes from time to time over multiple documents. Therefore summarization-visualization-tracking is necessary for any real life application. Researchers have tried with various types of output formats like textual or visual summary or overall tracking along time dimension. Several research attempts can be found in the literature on **Topic-wise** and **Polarity-wise** summarization and on **Visualization** and **Tracking**. The key issue regarding the sentiment aggregation is "***How the data should be aggregated, Topic-wise, Sentiment-wise or Otherwise?***". The experiments started with the multi-document topic-opinion textual summary. The 5W constituent based textual summarization-visualization-tracking system has been devised to meet the need for an *at-a-glance* presentation. The 5W constituent based aggregation system is a multi-genre system. The system facilitates users to generate sentiment tracking with textual summary and sentiment polarity wise graph based on any dimension or combination of dimensions as they want, for example, "*Who*" are the actors and "*What*" are their sentiment regarding any topic, changes in sentiment during "*When*" and "*Where*" and the reasons for change in sentiment as "*Why*". The 5W constituent based summarization-visualization-tracking system falls into every genre and attempts to answer the philosophical question "*Topic-Wise*, *Polarity-Wise* or *Other-Wise*".

The conclusion chapter of the thesis gives a summary of the experiments carried out and focuses on the new ideas put forward in the present work. It gives an account of the key contributions of the thesis and concludes by providing future possible avenues of this work. The key research contributions of the present work have been noted corresponding to each sub-problem in the area of sentiment analysis: Sentiment Lexicon Acquisition, Sentiment / Subjectivity Detection, Sentiment Polarity Detection, Sentiment Structurization, Sentiment Summarization-Visualization-Tracking.

Resource acquisition is one of the most challenging obstacles while working with resource constrained languages like Bengali. Extensive NLP research activities in Bengali have started recently but resources like annotated corpus, various linguistic tools are still unavailable for Bengali in the required measure. Corpus developments for subjectivity, polarity, structurization and summarization-visualization-tracking tasks have been discussed in the respective chapters. In this Appendix section, the development of the two main NLP tools in Bengali, i.e., **Stemmer** and **Dependency Parser** have been discussed. These NLP tools were not available when the work started and the development of these tools was taken up as an extension to the planned work for the thesis.

Thesis Title

Opinion Extraction and Summarization from Text Documents in Bengali

TABLE OF CONTENTS

4 Sentiment Structurization

LIST OF TABLES

LIST OF FIGURES

Introduction

Opinion Extraction and Summarization from Text Documents in Bengali

Sentiment analysis or opinion mining refers to the application of natural language processing, computational linguistics and text analytics to identify and extract sentimental/opinionated/emotional information from text. Actually sentiment analysis, opinion mining and emotion analysis refers to the same task. A basic task in sentiment analysis is the classification of the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed sentiment in a document, a sentence or an entity feature/aspect is positive (*happy*), negative (*sad*) or neutral (*memorable*). The feature/aspect level analysis of sentiment demands proper structures for more precise sentiment extraction. Sentiment/opinion aggregation is a necessary requirement from the end users' perspective. Therefore, summarization is necessary to present an at-a-glance presentation of the main points made in a single sentiment/opinion or how sentiment/opinion changes from time to time over multiple documents.

The works carried out as part of the thesis titled "***Opinion Extraction and Summarization from Text Documents in Bengali***" has been reported. A very brief idea about the motivation and the background necessity, which initiated the whole sentiment analysis research, has been reported in section I.1. Sentiment analysis (SA) from natural language text can be defined in terms of several sub-problems which have been discussed in section I.2. The overview of the present thesis has been reported in section I.3. Each sub-problem has been discussed in each chapter in the present thesis. Related works / literature survey for each sub-problem domain has been reported in that particular chapter.

Human sentiment knowledge grows with its age by day to day cognitive interactions. Sentiment is not a direct property of languages. An intelligent system should need some prior knowledge to act properly. Sentiment knowledge is generally wrapped into computational lexicon, technically called **Sentiment Lexicon.** The overall sentiment knowledge acquisition tasks in the present work have been reported in Chapter One of the thesis and have been introduced in sub-section I.3.1 Similar to classical pattern recognition problems, Sentiment Analysis is also classified into the identification and the classification genre called **sentiment / subjectivity detection** and **polarity classification** respectively. The proposed techniques for subjectivity detection and polarity classification are reported in the chapter Two and Chapter Three of the thesis respectively and have been introduced in sub-section I.3.2 and sub-section I.3.3 respectively. The need of the end user is the driving force behind the sentiment analysis research. The end users are not only looking for the binary (positive/negative) or multi-class sentiment classification but they are more interested in aspectual/structural sentiment analysis. Therefore only sentiment detection and classification is not enough to satisfy the need of the end user. Structurization involves identification of various aspects of a sentiment/opinion, i.e., sentiment holder, sentiment topic, domain dependant attributes and so on. Proper **structurization** of sentiments is essential before proceeding for any further granular analysis or generation and aggregation. The whole research attempts on structurization are described in the Chapter Four of the thesis and have been briefly introduced in the sub-section I.3.4 To meet the satisfaction level of end users' an intelligent sentimental/opinionated information processing system should be capable enough to present an *at-a-glance* presentation of aggregated information, scattered over various sources / documents. Finally,

textual or visual **summarization, visualization** or **tracking** of sentiment are the striking needs from the perspective of the end user. The overall summarization-visualization-tracking research attempts are described in the Chapter Five of the thesis and have been introduced in the sub-section I.3.5. The various experiments described in the present thesis mainly deals with **English** or **Bengali** languages or both.

The Conclusion chapter of the thesis summarizes the works carried out as part of the present work, identifies the key scientific contributions of the works and concludes by providing the future roadmap. These issues have been introduced in sub-section I.3.6. Resource acquisition and appropriate tool development are the most challenging obstacles while working with resource constrained languages like Bengali. The corpus developments/annotations, other resource acquisition processes and appropriate tool development for subjectivity detection, polarity classification, structurization and summarization-visualization-tracking tasks have been discussed in the respective chapters. The development of the two main tools, i.e. **Stemmer** and **Dependency Parser** has been described in the Appendix section of the thesis while these have been introduced in sub-section I.3.7.

I.1 Why Sentiment Analysis / Opinion Mining?

Necessity is the mother of all invention

---Plato

Sentiment Analysis/Opinion Mining is one of the most pursued research topics in recent times. Recently, many researchers and companies have explored the area of opinion detection and analysis. With the increased number of Internet users, there is a proliferation of opinions available on the web. Not only do we read more opinions from the web, such as in daily news editorials, but also we post more opinions through mechanisms such as governmental web sites, product review sites, news group message boards, personal blogs and twitters. This phenomenon has opened the door for massive opinion collection, which has potential impact on various applications such as public opinion monitoring and product review summary systems.

Moreover in today's digital age, text is the primary medium of representing and communicating information, as evidenced by the pervasiveness of e-mails, instant messages, documents, weblogs, news articles, homepages and printed materials. Our lives are now saturated with textual information and there is an increasing urgency to develop technologies to help us manage and make sense of the resulting information overload.

While expert systems have enjoyed some success in assisting information retrieval, data mining and natural language processing (NLP) systems, there is a growing necessity of sentiment analysis systems that can automatically process the plethora of sentimental information available in online electronic

text. The increasing social necessity is the driving force for the massive research effort on Sentiment Analysis/Opinion Mining.

I.2 What is Sentiment Analysis / Opinion Mining?

Any scientific research needs to know the proper definitions of the problems in order to solve it. The essential question that is raised at the beginning of the sentiment analysis research is "***What is sentiment or opinion?***". Various research endeavors attempted to answer this question in the light of psychology, philosophy, psycholinguistics and even cognitive science. The researchers attempted to give their own definitions. Among those research endeavors, the General Inquires (Stone, 1966) System and the Subjectivity definition by Janyce Wiebe (Wiebe et. al., 1990) are the milestones that mark the avenue to the current research trend of today.

The sentiment analysis research started as a content analysis research problem in the behavioral science. The General Inquirer System[1] (Stone, 1966) is the first attempt in this direction. The aim was to gain understanding of the psychological forces and perceived demands of the situation that were in effect when the document was written. The system usually counts the occurrences of positive or negative emotion instances in any particular piece of text. Although the sentiment analysis research has started long back but still the question "***What is sentiment or opinion?***" remain unanswered till date! It is very hard to define sentiment or opinion and to identify the regulating or the controlling factors of sentiment. Moreover no concise set of psychological forces could be defined that really affect the writers' sentiments, i.e., broadly the human sentiment.

"*How the mind works is still a mystery. We understand the hardware, but we don't have a clue about the operating system.*"

--James Watson (Nobel laureate)

"Opinion is the Medium between Knowledge and Ignorance."

--Plato

Probably the question may not be answered by the theories of the computer science and may be the scopes of the medical science, Cognitive Science or Psychologies have to be explored. The ***Topical Relevant Opinionated Sentiment*** detection is well known a***s Subjectivity Detection*** (Wiebe et. al., 1990)***.*** Janyce Wiebe borrowed the definition of opinion from a Psycholinguistics research which states: ***an opinion could be defined as a private state that is not open to objective observation or verification*** (Quirk et. al., 1985).

[1] http://www.wjh.harvard.edu/~inquirer/

Sentiment Analysis/Opinion Mining from natural language text is a multifaceted and multidisciplinary AI problem. It tries to narrow the communicative gap between the highly sentimental human and the sentimentally challenged computers by developing computational systems that can recognize and respond to the sentimental states of the human users. There is a perpetual debate about the better ways of collecting intelligence either by following the functional path of biological human intelligence or generating new methodologies for completely heterogeneous mechatronics machine and redefine a completely new horizon called electronic intelligence. The research endeavors in the present task is to find out the optimum solution strategies for machines that either mimic the techniques of self-organized biological human intelligence or at least can simulate the functional similarities of human sentimental intelligence.

I.3 Overview of the Present Thesis

The present thesis is distributed in five chapters. The first chapter describes the Sentiment knowledge acquisition process in terms of **Sentiment Lexicon** while the second and third chapter describes the **subjectivity detection** and **polarity classification** problem respectively. To produce a formidable output for the end user proper structurization is required. The **structurization** is described in the chapter four and finally the **Summarization-Visualization-Tracking** methodologies are described in the chapter five. Finally in the conclusion section, the key contributions of the present work in the Sentiment Analysis research have been self explained. In the appendix section, the details of the natural language processing tools, Dependency Parser and Stemmer for Bengali have been reported. The following subsections give a very brief description about the chapters in the present thesis.

I.3.1 Sentiment Knowledge Acquisition (Chapter 1)

Sentiment knowledge acquisition in terms of sentiment lexicon is the vital pre-requisite of any sentiment analysis system. Previous studies have proposed to attach **prior polarity** (Esuli and Sebastiani, 2006) to each sentiment lexicon level. Prior polarities are approximate values and are based on corpus statistics.

A number of research endeavors can be found in the literature for creation of sentiment lexicon in several languages and domains. These techniques can be broadly categorized into two genres, one follows the classical manual annotation techniques (Andreevskaia and Bergler, 2006; Wiebe and Riloff, 2005; Mohammad et. al., 2008) and the other includes various automatic techniques (Tong, 2001; Mohammad and Turney, 2010). Both types of techniques have few limitations. Manual annotation techniques are undoubtedly trustable but it generally takes time for development. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain. Manual annotation techniques require a large number of annotators to balance the

sentimentality of individual annotators in order to reach agreement. But qualified human annotators are quite unavailable and also costly.

Both the processes have been attempted to develop SentiWordNet(s) (Das and Bandyopadhyay, 2010(e)) for multiple languages. During evaluation, it has been observed that there are two issues that should be satisfied by a good quality sentiment lexicon. The first one is the **coverage** and the second one is the **credibility** of the associative polarity scores. It may be concluded that automatic processes are good for coverage expansion but manual methods are trustable for prior polarity assignment.

The automatic processes used in the present work are ***bilingual dictionary based approach, WordNet based synonym and antonym expansion, orthographic antonym generation*** and ***corpus based approach***. English sentiment lexicons have been chosen as the source and then the synset members have been translated into the target language using bilingual dictionaries. WordNet 3.0 has been effectively used to expand a given synset via synonym or antonym search. Sixteen hand crafted suffix/affix rules have been used to orthographically create more antonyms for a given synset and corpus validation have been done later to confirm the validity of the orthographically generated form. The generated sentiment lexicon has been used as a seed list. Language specific corpus has been automatically tagged with these seed words using the simple tagset of SWP (Sentiment Word Positive) and SWN (Sentiment Word Negative). A Conditional Random Field (CRF) based classifier has been trained on the tagged corpus and then applied on un-annotated corpus to find out new language and culture specific sentimental words. These techniques have been successfully used for three Indian languages: **Bengali, Hindi** and **Telugu** (Das and Bandyopadhyay, 2010(c); Das and Bandyopadhyay, 2010(e)). The Bengali SentiWordNet[2] (Das and Bandyopadhyay, 2010 (f)) have already been made available for further research.

As there is scarcity of human annotators, it has been decided to involve the **Internet Population** for creating more credible sentiment lexicons (Das and Bandyopadhyay, 2011; Das, 2011). Internet population is very huge in number and is ever growing (approximately, 360,985,492)[3]. It includes people with various languages, cultures, age etc. and thus it is not biased towards any domain, language or particular society. An online game called **Dr Sentiment** has been developed which is a template based interactive online game. **Dr Sentiment** collects players' sentiment by asking a set of simple template based questions and finally reveals sentimental status of the player. The lexicons tagged by this system are credible as it is tagged by human beings. It is not a static sentiment lexicon set as the prior polarity scores are updated regularly. Almost 100 players per day are currently playing it throughout the world in different languages. **Global SentiWordNet** (Das and Bandyopadhyay, 2010(d)), the SentiWordNet(s) for **57 languages,** has been developed using Google translation API services[4].

[2] http://www.amitavadas.com/sentiwordnet.php
[3] http://www.internetworldstats.com/stats.htm
[4] http://translate.google.com/

Dr Sentiment also helps to capture an overall picture of human social psychology regarding sentiment understanding. Figure I.1 and Figure I.2 show how overall sentimentality changes with age and gender respectively. Figure I.3 shows how sentimentality changes with geospatial locations. The word "blue" gets tagged by different players around the world. But surprisingly it has been tagged as positive from one part of the world and negative from another part of the world.

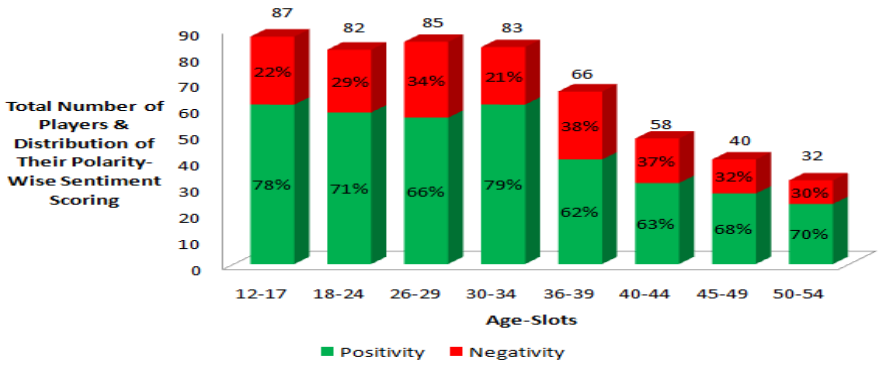


Figure I.1: Senti-Mentality Age Wise

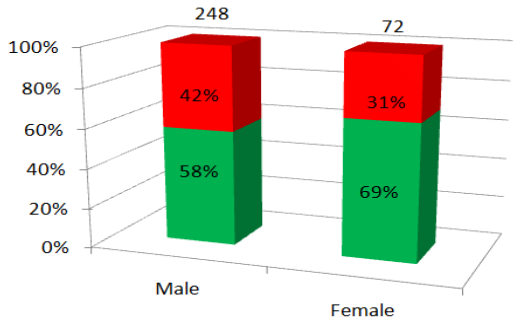


Figure I.2: Senti-Mentality Gender Wise

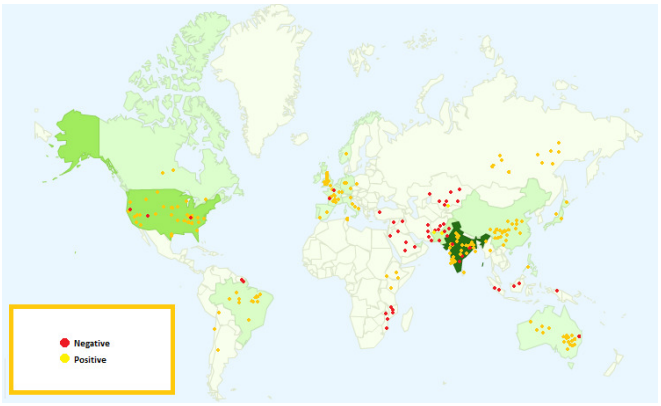


Figure I.3: Geospatial Senti-Mentality

Presently, several psychological information are being incorporated into the existing SentiWordNet and the resultant lexicon is termed as the **PsychoSentiWordNet** (Das, 2011). The PsychoSentiWordNet holds variable prior polarity scores that may be fetched depending upon the regulating psychological aspects. The following example illustrates the definition better.

Aspect values (Profession)	Input	Polarity
Null	High	Positive
Businessman	High	Negative
Share Broker	High	Positive

I.3.2 Sentiment / Subjectivity Detection (Chapter 2)

The term subjectivity simply refers to the identification of sentiments in a piece of text. More precisely, the term ***Subjectivity*** can be defined as the ***Topical Relevant Opinionated Sentiment*** (Wiebe et. al., 1990)**.** The subjectivity is concerned with whether the expressed sentiment is related to the relevant topic or it fulfills the overall desired goal of a Sentiment Analysis system.

Sentiment or subjectivity detection is a very tough challenge for emotionally challenged machines and even for human beings. Let us take a look at the following example.

Example 1: **My camera broke in two days.**

Example 2: Type: Film Review, Film Name: Deep Blue Sea, Holder: Arbitrary-outside of theatre

This is blue!

In the first example, it is very hard to disambiguate whether the author is only talking about an accident or complaining about the quality of the camera. The problem with the second example is that there is no evaluative expression and no clue at syntax or semantics level to identify the sentiment.

Previous studied have already identified some clues at the lexical and the syntactic level (Aue and Gamon, 2005; Hatzivassiloglou and McKeown, 1997; Nasukawa and Yi, 2003). A series of experiments have been carried out to find out the best and optimum feature set for English and Bengali language. The final feature set used for the experiments have been classified into three genres as reported in the Table I.1.

On the algorithmic aspect, the experiments started with the **Rule-based** (Das and Bandyopadhyay, 2009(c)) technique and continued with **Machine Learning** (Das and Bandyopadhyay, 2009(b)) and **Hybrid** techniques (Das and Bandyopadhyay, 2009(a)). The Theme Detection technique has been developed to detect topical relevant sentiments. The themes relate to the sentimental topic of any document. But there may be more unrevealed clues based on human psychology or on complex

relationships among the linguistic clues for sentiment / subjectivity detection which may not be extracted with present NLP/simple machine learning techniques.

Types	Features
Lexico-Syntactic	POS
	SentiWordNet
	Frequency
	Stemming
Syntactic	Chunk Label
	Dependency Parsing
Discourse Level	Title of the Document
	First Paragraph
	Average Distribution
	Theme Word

Table I.1: Features for Subjectivity Detection

Thus, experiments have been carried out with **Genetic Algorithm** (Das and Bandyopadhyay, 2010(g)) to adopt the biological evolutionary path of the human intelligence for machines. The accuracy of the system with Genetic-Based Machine Learning (GBML) technique reaches **90.22%** (MPQA: news) and 93.00**%** (IMDB: movie review) for English and **87.65%** (news) and **90.6%** (blog) for Bengali as stated in the Table I.2.

Languages	Domain	Precision	Recall
English	MPQA	90.22%	96.01%
	IMDB	93.00%	98.55%
Bengali	NEWS	87.65%	89.06%
	BLOG	90.6%	92.40%

Table I.2: Results of Final GA based Subjectivity Classifier

Machine learning algorithms in NLP systems generally experiment with various combinations of syntactic and semantic linguistic features to identify the most effective feature set. The sentiment / subjectivity detection problem in the present task has been viewed as a **Multi-Objective** or **Multi-Criteria Optimization** search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The **fitness function** calculates the accuracy of the subjectivity classifier based on the feature set identified by **natural selection** through the process of **crossover** and **mutation** after each generation. The GBML technique automatically identifies the best feature set based on the principle of natural selection and **survival of**

the fittest. The identified fittest feature set is then optimized **locally** and **global optimization** is obtained by multi-objective optimization technique.

I.3.3 Sentiment Polarity Detection (Chapter 3)

The polarity classification task involves sentiment/opinion classification into semantic classes (Turney et. al., 2002) such as *positive, negative or neutral* and/or other fine-grained emotion classes like *happy, sad, anger, disgust and surprise.*

The two step methodology, i.e., use of prior polarity lexicon followed by any NLP technique is the standard method for the polarity classification task as established by several previous research efforts (Hatzivassiloglou et. al., 1997; Strapparava and Mihalcea, 2008; Denecke, 2009). The SentiWordNet (Bengali) (discussed in the Chapter One) developed as part of the present work has been used as the prior polarity lexicon. The application of the NLP techniques started with the Syntactic-Statistical classification technique (Das and Bandyopadhyay, 2010(a); Das and Bandyopadhyay, 2010(h)). The syntactic clue directly helps to understand the relation between the localized semantic orientations, i.e., word level semantic orientation and the contextual semantic orientation which is basically the word/phrase/sentence level semantic orientation. In the following example sentence, the localized semantic orientation at word level: ভালো (good) can be obtained directly from the prior polarity lexicon.

He is not a good^{+} boy.

সে ভালো$^{+}$ ছেলে নয়।

But the negation word changes the contextual semantics in the opposite direction, i.e., it turns the semantics to negative. In the sentence, the word "not (নয়)" has a modifier relationship with the word "good (ভালো)" (modified). Therefore, it is very easy to infer that the resultant contextual semantic orientation of the sentence is negative. Sometimes, the syntax helps to predict the semantic orintation of any new word. Let us consider the following sentence.

This is ugly^{-} and smelly.

এটি বিশ্রী$^{-}$ এবং কটুগন্ধযুক্ত।

Let us further consider that the prior polarity dictionary only covers the word "*ugly (বিশ্রী)"* with negative semantic orientation and not the word *"smelly (কটুগন্ধযুক্ত)"*. It is more or less obvious that the semantic orientation of the new word *"smelly (কটুগন্ধযুক্ত)"*will be negative because it has been observed that words with same orientation are in syntactic conjunction with "and" and words with orthogonal semantic orientation are in syntactic disjunction with "but/rather/either…etc". Let us consider the following example.

Good^{+} but costly^{-}.

ভালো[+] কিন্তু দামী[-]।

It has also been observed that localized syntax helps to understand the discourse level sentimental semantics to some extent (Somasundaran, 2010). For example, suppose the following sentences are from two different paragraphs from the same document.

The reason behind the electoral disaster is the wrong (strategy of the previous Government).

পূর্বতন সরকারের ভুল নীতি ভোটে ভরাডুবির অন্যতম কারন।

We will not follow the (strategy of the previous government), said Mamata Banerjee.

মমতা ব্যানার্জী বলেন আমরা পূর্বতন সরকারের নীতি অনুসরন করবো না।

In the first sentence, the word "*wrong (ভুল)*" is modifying the phrase "*strategy of the previous Government (পূর্বতন সরকারের নীতি)*" and it is negative. Therefore in the same scope of the document it is very likely that a single author will not sentimentally differ too much regarding the same topic. Thus, the final semantic orientation of the second sentence is likely to be positive. But it is very hard to incorporate this kind of knowledge into the Syntactic-Statistical polarity classifier. An in-depth semantic tagging is required for this kind of work. The syntactic statistical polarity classifier in the present work starts with phrase/sentence level polarity classification.

The Support Vector Machine (SVM) has been used with a number of features for the development of the syntactic polarity classifier. The polarity classification task mainly involves **syntactic analysis** like **Modifier-Modified** relationship or **Association ambiguity**. A **dependency parser** (Ghosh et. al., 2009(d); Ghosh et. al., 2010(m)) has been developed for Bengali language as there was no dependency parser available. The dependency parser helps to properly analyze the syntactic structure of the language. The detailed development process of the Dependency Parser is described in the Appendix section. Many other linguistics features have been included in the polarity classifier. **Feature ablation** method has shown that only the dictionary based method gives a good **baseline** whereas other features are needed to disambiguate the syntactic nature of any language (Das and Bandyopadhyay, 2010(a));(Das and Bandyopadhyay, 2010(h)). The results have been reported in Table I.3. The Sentiment polarity classifier achieves its best performance of 70.04% with the negative word, stemming cluster, functional word, parts of speech, chunk and dependency tree features along with the SentiWordNet information.

Dealing with unknown/new words is a common challenge for NLP systems. It becomes more difficult for sentiment analysis because it is very hard to find out any contextual clue to predict the sentimental orientation of any unknown/new word. A prior polarity lexicon is attached with two probabilistic values, i.e., positivity and negativity scores and there is no clue in the SentiWordNet regarding "***which value to pick in what context?"***. The general trend is to pick the highest one but that may vary depending on the

context. For example, the word "***High***" (Positivity: 0.25, Negativity: 0.125 for "***High***" from the SentiWordNet) is attached with a positive (positivity value is higher than the negativity value) polarity in the sentiment lexicon but the polarity of the word may vary as it happens in the following example sentences. In the first sentence, the word ***"High"*** has a positive polarity while in the second sentence the polarity is negative.

Sensex reaches high^{+}.

Price goes high^{-}.

Features	Accuracy
SentiWordNet	47.60%
SentiWordNet + Negative Word	50.40%
SentiWordNet + Negative Word + Stemming Cluster	56.02%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word	58.23%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word+ Parts Of Speech	61.9%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech +Chunk	66.8%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech + Chunk +Dependency tree feature	70.04%

Table I.3: Performance of the Syntactic Polarity Classifier by Feature Ablation

Additional NLP techniques are required to disambiguate these types of words. There are 6619 lexicon entries in the English SentiWordNet where both the positivity and the negativity values are greater than zero. Similarly, there are a total of 17927 lexical entries in the English SentiWordNet, whose positivity and negativity value difference is less than 0.2. These statistics are reported in the Table I.4. These entries are ambiguous because there is no clue in the SentiWordNet regarding the positivity or negativity of such entries.

Type	Number
Total Token	**115424**
Positivity>0 && Negativity>0	**6619**
Positivity>0 || Negativity>0	**28430**
Positivity>0 && Negativity=0	**10484**
Positivity=0 & Negativity>0	**11327**
$\lvert Positivity - Negativity \rvert$ >=0.2	**17927**

Table I.4: Ambiguous Entries in SentiWordNet

The research attempts in the present work are mainly concerned about the ambiguous entries from the SentiWordNet. The basic hypothesis is that if some sort of contextual information can be added in the sentiment lexicon along with the prior polarity scores then the updated rich lexicon network will serve better than the existing one and it may lessen the requirements for further NLP techniques to disambiguate the contextual polarity. A new paradigm called ***Sentimantics*** has been introduced in the present work which uses distributed Semantic Lexical Models to hold the sentiment knowledge with contextual common sense (Das and Bandyopadhyay, 2012(c)). Such a paradigm has not been explored before.

Two different type models for Sentimantic composition have been examined that are empirically grounded and can represent the contextual similarity relations among various lexical sentiment and non-sentiment concepts. The experiments started with existing resources like ConceptNet and SentiWordNet for English and SemanticNet (Das and Bandyopadhyay, 2010(p));(Das, 2010(n)) and SentiWordNet (Bengali) for Bengali. The common sense lexicons like ConceptNet and SemanticNet are developed for general purpose and the formalization of Sentimantics from these resources is challenging due to lack of dimensionality. A Vector Space Model (VSM) has been developed by a corpus driven semi-supervised method to hold the Sentimantics from scratch. This model performs relatively better than the previous one and is quite satisfactory.

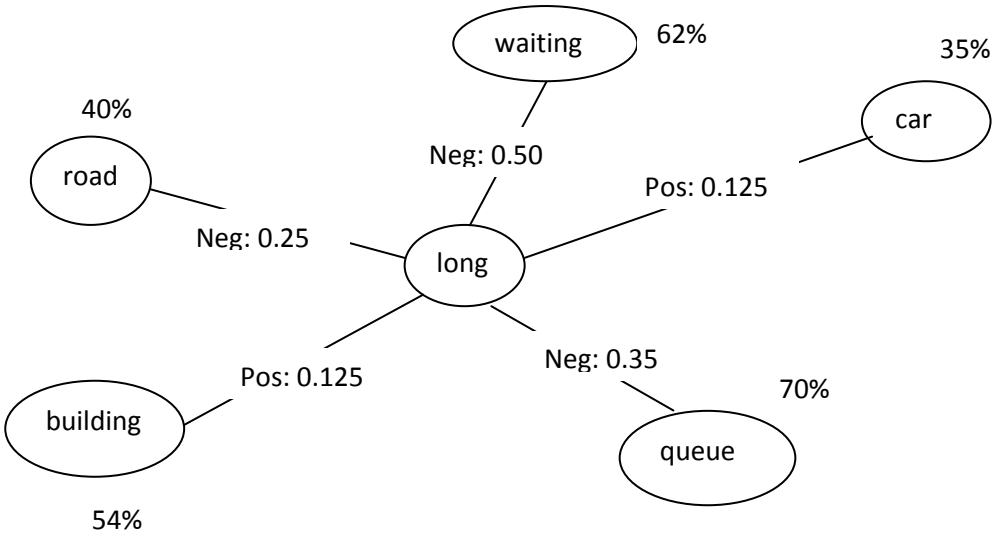


Figure I.4: The developed Sentimantics Network by Network Overlap Technique

Generally extracting knowledge from this kind of VSM is algorithmically very expensive because it is a very high dimensional network. Another important limitation of this type of model is that it demands very well defined processed input to extract knowledge, like, Input: (high) Context (sensex, share market, point), which demands NLP pre-processing steps on the input text to extract knowledge from this VSM. Finally, the Syntactic Co-Occurrence Based VSM with relatively fewer dimensions has been proposed. The final model is the best performing lexicon network model and may be described as the

acceptable solution for the Sentimantics problem. The details of the proposed models are described in the Chapter Three.

Each sentiment word in the developed lexical network by the Network overlap technique is assigned a contextual prior polarity. Figure I.4 shows the lexical network for the word "*long*".

I.3.4 Sentiment Structurization (Chapter 4)

The need of the end user is the driving force behind the sentiment analysis research. The outcomes of these research endeavors should lead to the development of a real time sentiment analysis system, which will successfully satisfy the need of the end users. Let us have a look at some real life needs of the end user. For example, a market surveyor from company A may identify the need to find out the changes in public opinion about their product X after release of product Y by another company B. The different aspects of product Y that the public consider better than product X are also points of interest. These aspects could typically be the durability of the product, power options, weight, color and many more other issues that depend on the particular product. In another scenario, a voter may be interested to study the change of public opinion about any leader or any public event before and after any election. In this case the aspect could be a social event, economic recession and may be other issues. The end users are not only looking for the binary (positive/negative) sentiment classification but they are more interested in aspectual sentiment analysis. Therefore, only sentiment detection and classification is not enough to satisfy the need of the end user. A sentiment analysis system should be capable enough to understand and extract out the aspectual sentiments present in a natural language text.

Previous research efforts have already proposed various structures or components for sentiment extraction. Among the proposed sentiment structures the most widely used structures are **Holder** (Kim and Hovy, 2004**;** Choi et. al., 2005; Bethard et. al., 2006), **Topic** (Ku et. al., 2005; Zhou et. al., 2006; Kawai et. al., 2007) and other domain dependent **Attributes** (Kobayashi et. al., 2006; Bal and Saint-Dizier, 2009). But the real life users are not always interested about all the aspects at a time, rather they look for opinion/sentiment changes of any "Who" during "When" and depending upon "What" or "Where" and "Why". With this hypothesis, the 5W (**Who**, **What**, **When**, **Where** and **Why**) constituent (Das et. al, 2010(i)) extraction technique for sentiment/opinion structurization has been proposed. The proposed 5W structure is domain independent and more generic than the existing semantic constituent extraction structure.

Table I.5 presents the sentence level co-occurrence patterns of the 5Ws in the corpus. 5Ws do not appear together regularly in the corpus. Hence sequence labeling with 5W tags using any machine learning technique will lead to a label bias problem and may not be an acceptable solution for the present problem of 5W role labeling. Therefore, a system that follows hybrid architecture has been proposed. It statistically assigns 5W labels to each chunk in a sentence using the Maximum Entropy Model (MEMM). A rule based post-processor helps to reduce many false hits by the MEMM based system and at the same time also identifies new 5W labels. The rules have been developed based on the

acquired statistics on the training set and the linguistic analysis of standard Bengali grammar. By analyzing the output of both the MEMM and the hybrid systems (MEMM followed by the rule based post-processor system) it can be easily inferred that the hybrid structure is essential for this 5W problem domain.

Tags	Percentage					
	Who	What	When	Where	Why	Overall
Who	-	58.56%	73.34%	78.01%	28.33%	73.50%
What	58.56%	-	62.89%	70.63%	64.91%	64.23%
When	73.34%	62.89%	-	48.63%	23.66%	57.23%
Where	78.0%	70.63%	48.63%	-	12.02%	68.65%
Why	28.33%	64.91%	23.66%	12.02%	-	32.00%

Table I.5: Sentence Level Co-occurrence Pattern of 5Ws

I.3.5 Sentiment Summarization-Visualization-Tracking (Chapter 5)

Aggregation of information is the necessity from the end users' perspective but it is nearly impossible to develop consensus on the output format or how the data should be aggregated. Researchers have tried with various types of output formats like textual or visual summary or overall tracking along time dimension. Several research attempts can be found in the literature on **Topic-wise** (Yi et. al., 2003; Pang and Lee, 2004; Zhou et. al., 2006) and **Polarity-wise** (Hu, 2004; Yi and Niblack, 2005; Das and Chen, 2007) summarization and on **Visualization** (Morinaga et. al., 2002; Gamon et. al., 2005; Carenini et. al., 2006) and **Tracking** (Lloyd et. al., 2005**;** Mishne and Rijke, 2006; Fukuhara et al., 2007). The key issue regarding the sentiment aggregation is "***how the data should be aggregated?***". Dasgupta and Ng (Dasgupta and Ng, 2006) throw an important question: "***Topic-wise, Sentiment-wise or Otherwise?***" about the opinion summary generation techniques. Actually the output format varies on end users' requirements and the domain. Several output formats have been experimented in the present work.

The experiments started with the multi-document topic-opinion textual (Das and Bandyopadhyay, 2010(j));(Das and Bandyopadhyay, 2010(k)) summary. The 5W constituent based textual summarization-visualization-tracking (Das and Bandyopadhyay, 2012(d)) system has been devised to meet the need for an *at-a-glance* presentation. The 5W constituent based aggregation system is a multi-genre system. The system facilitates users to generate sentiment tracking with textual summary and sentiment polarity wise graph based on any dimension or combination of dimensions as they want, for example, "*Who*" are the actors and "*What*" are their sentiment regarding any topic, changes in sentiment during "*When*" and "*Where*" and the reasons for change in sentiment as "*Why*". The 5W constituent based summarization-visualization-tracking system falls into every genre and attempts to answer the philosophical question "*Topic-Wise*, *Polarity-Wise* or *Other-Wise*".

Topic-Wise: The system facilitates users to generate sentiment summary based on any customized topic like Who, What, When, Where and Why based on any dimension or combination of dimensions as they want.

Polarity-Wise: The system produces an overall gnat chart that can be treated as the overall polarity wise summary. An interested user can still look into the summary text to find out more details.

Visualization and Tracking: The system facilitates users to generate visual sentiment tracking with polarity wise graph based on any dimension or combination of dimensions as they want, i.e., "Who" are the actors and "What" are their sentiment regarding any topic, changes in sentiment during "When" and "Where" and the reasons for change in sentiment as "Why". The final graph for tracking is generated with a timeline.

Moreover, the end user can structure their information need as:

- ***Who***? Who was involved?
- ***What***? What happened?
- ***When***? When did it take place?
- ***Where***? Where did it take place?
- ***Why***? Why did it happen?

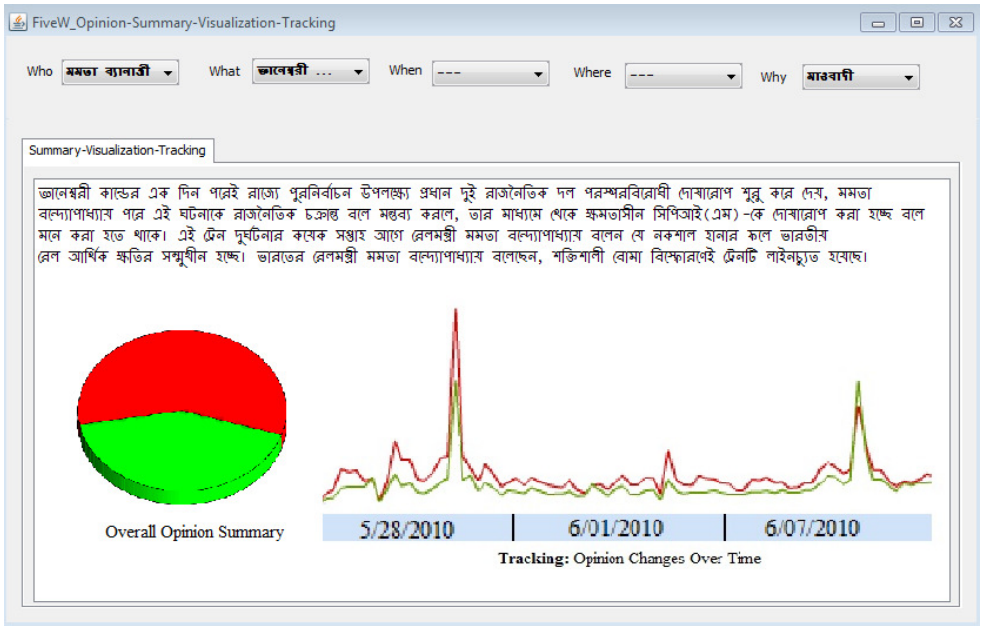


Figure I.5: Snapshot of the 5W Sentiment Summarization-Visualization-Tracking System

During the development of the multi-document topic-opinion summarization system, a strong semantic lexical network (Das and Bandyopadhyay, 2010(j); Das and Bandyopadhyay, 2010(k)) has been proposed following the idea of Mental Lexicon models. The same lexical semantic network has been used to develop the 5W system.

The present 5W summarization-visualization-tracking system (Das and Bandyopadhyay, 2012(d)) facilitates users to generate sentiment summary and sentiment polarity wise graph based visualization on any dimension and combination of dimensions as desired by the user. The present system also provides an overall summary. A snapshot of the 5W Sentiment Summarization – Visualization – Tracking system has been presented in Figure I.5. Another important aspect of the present system is that a user can provide no input along any dimension to see the all the possible information on that dimension.

I.3.6 Conclusion

The conclusion chapter of the thesis gives a summary of the experiments carried out and new ideas put forward in the present work. It gives an account of the key contributions of the thesis and concludes by providing future possible avenues of this work. The key research contributions of the present work have been noted corresponding to each sub-problem in the area of sentiment analysis: Sentiment Lexicon Acquisition, Sentiment / Subjectivity Detection, Sentiment Polarity Detection, Sentiment Structurization, Sentiment Summarization-Visualization-Tracking.

I.3.7 Appendix: Stemmer and Dependency Parser

Resource acquisition is one of the most challenging obstacles while working with resource constrained languages like Bengali. Extensive NLP research activities in Bengali have started recently but resources like annotated corpus, various linguistic tools are still unavailable for Bengali in the required measure. Corpus developments for subjectivity, polarity, structurization and summarization-visualization-tracking tasks have been discussed in the respective chapters. In this Appendix section, the development of the two main NLP tools in Bengali, i.e., **Stemmer** (Das and Bandyopadhyay, 2010(l)) and **Dependency Parser** (Ghosh et. al., 2009(d); Ghosh et. al., 2010(m)) have been discussed. These NLP tools were not available when the work started and the development of these tools was taken up as an extension to the planned work for the thesis.

Chapter 1

Sentiment Knowledge Acquisition

Lexical analysis plays a crucial role to identify sentiments/opinion from a text. For example, words like ***love***, ***hate***, ***good*** and ***favorite*** directly indicate sentiment. As sentiment is a property of human intelligence and is not entirely based on the features of a language, thus prior non-linguistic knowledge is required for automatic sentiment analysis. Sentiment knowledge acquisition in terms of sentiment lexicon is a vital pre-requisite of any sentiment analysis system. Previous studies have proposed to attach **prior polarity** to each sentiment lexicon level. Prior polarity is an approximation value based on statistics collected from corpus.

A number of research endeavors could be found in the literature for creation of sentiment lexicon in several languages and domains. These techniques can be broadly categorized into two genres, the first one follows the classical manual annotation techniques (Andreevskaia and Bergler, 2006; Wiebe and Riloff, 2005; Mohammad et. al., 2008) and the other includes various automatic techniques (Tong, 2001; Mohammad and Turney, 2010). Both types of techniques have few limitations. Manual annotation techniques are undoubtedly trustable but it generally takes time. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain. Manual annotation techniques require a large number of annotators to balance one's sentimentality in order to reach agreement. But human annotators are quite unavailable and costly.

Both the automatic and manual processes have been attempted to develop SentiWordNet(s) (Das and Bandyopadhyay, 2010(e)) for multiple languages. During evaluation it has been noticed that there are two issues should be satisfied to be a good qualitative sentiment lexicon. The first one is **coverage** and the second one is **credibility** of the associative polarity score. The experimentation started with Bengali[1] (ethnonym: Bangla; exonym: Bengali) language. Bengali is a computationally resource scarce language but it is the fifth popular language round the globe, second in India and the national language in Bangladesh. Techniques have been developed for cross lingual projection of Bengali sentiment lexicon from English as a source language. Later on it has been showed how these techniques can be replicated for other Indian languages (Hindi and Telugu) as well. Finally, Dr. Sentiment, a template based online interactive gaming technology, has been proposed to automatically collect the lexical level sentiment polarity involving Internet population. This technique excels over all the existing methodologies and finally sentiment lexicons are being created for 57 international languages.

At first, the concept of standard prior polarity lexicon has been elaborated along with previous studies for sentiment knowledge acquisition. In the subsequent sections the research endeavors in the present work for multi-lingual sentiment knowledge acquisition have been elaborated.

1.1 Prior Polarity Sentiment Lexicon

The current research trend is to attach prior polarity to each entry in the sentiment lexicon. Prior polarity is an approximate value based on statistics collected from corpus. The qualitative criterion for

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

any kind of lexical resource or dictionary is good coverage. It assures that the resource is useful across various domains. An additional issue for qualitative analysis of sentiment lexicon is the credibility of the attached prior polarity scores. Such credibility proves the authenticity and usability of the sentiment lexicon across various domains and possibly for various languages.

1.1.1 The Challenges in Prior Polarity Sentiment Lexicon

Polarity assignment of sentiment lexicons is a semantic disambiguation problem. The important aspects that govern the lexical level semantic orientation are natural language context (Pang et al., 2002), language properties (Wiebe and Mihalcea, 2006), domain pragmatic knowledge (Aue and Gamon, 2005), time dimension (Read, 2005), colors and culture (Strapparava and Ozbal, 2010) and many more unrevealed hidden aspects. The various aspects are now discussed in detail with appropriate examples.

1.1.1.1 Contextuality

I prefer Limuzin as it is **longer** than Mercedes.

Avoid **longer** baggage during excursion in Amazon.

In the previous two example sentences, the word "**longer**" has been used but the semantic orientation of the sentences is completely orthogonal. In the first sentence the word "**longer**" depicts positivity and explains why the author likes Limuzin. Limuzin is spacious than Mercedes and this information is wrapped in the word "**longer**". The second sentence expresses precaution. The hidden semantics of the sentence is that one should avoid longer baggage during excursion to a particular location as the **longer** baggage will be heavy to carry. During excursion people generally have to walk and heavy baggage should be avoided for this reason.

A similar example may be the word "**love**", which generally gives a sense of positivity but in the following example it poses a sense of negativity, more specially cruelty.

I **love** the way Prof. Moriarty kills!

The example sentence is extracted from a movie review. Prof. Moriarty is the famous rival character of Sherlock Holmes, the famous creation by Sir Arthur Conan Doyle.

The previous examples shows how lexical level polarity changes with context information. Therefore it is very difficult to assign a fixed polarity score to a particular lexicon entry.

1.1.1.2 Language-Culture Properties

সহেরা (Sahera: A marriage-wear of India)

দুর্গাপুজো (Durgapujo: A festival of Bengal, India)

The above words exhibit language-culture specific sentimentality. The following example sentences show how a marriage-wear or festival can trigger positive sentiment.

राहुल गांधीने अज पहली बार जितका **सेहरा** बंधा ।

(Today first time Rahul Gandhi wears the Sahera of his victory)

সামনে দূর্গাপূজো, চারিদিকে ছুটির আমেজ।

(**Durgapuja** is approaching and there is a mood of vacation everywhere)

The sentiment lexicon for any language should include such language and culture specific words to increase the coverage.

1.1.1.3 Domain Knowledge

Sensex goes **high**.

Price goes **high**.

In the previous two example sentences the word "**high**" expresses completely opposite semantic orientation. In the first sentence a sense of positivity is depicted whereas the sense is negative in the second sentence. We, human beings, understand this with our prior domain knowledge. It is thus very challenging to assign a fixed polarity at lexicon level. Another example word is "**unpredictable**". If it is said about a movie plot then probably it is positive as we all like unpredictable/interesting movies. But if the same word is used as "**unpredictable steering**" for a car then undoubtedly the sense is negative.

1.1.1.4 Time Dimension

Time is a vital issue to determine sentiment polarity of a word as people's sentiment changes along the time dimension. Let us consider the following text:

During 90's mobile phone users generally reported in various online reviews about their ***color-phones*** but in recent times ***color-phone*** is not just enough. People are fascinated and influenced by touch screen and various software(s) installation facilities on these new generation gadgets.

The case for **magnetic-tape** is quite similar as Compact Disks (CD) or Digital Video Disks (DVD) are the most recent technologies. This signifies that polarity scores need to be updated along the time dimension.

1.1.1.5 Colors and Culture

There is an in-depth relationship between color and sentiment. We frequently use colors in a text in order to express our emotion more vividly. Numerous experiments have been carried out in psycholinguistics, cognitive or medical science to understand the behavioral characteristics of colors and

how it affects our emotions. For instance, we usually stress the redness of someone's face to imply his/her anger or excitement, or we use phrases involving the black color to refer to a depressed mood. On the other hand, the pink color is mostly used with positive connotations such as 'to see everything in pink light', where the meaning is related to optimism and happiness. (Strapparava and Ozbal, 2010) have presented a nice experiment with natural language text and have compared their findings with previous psycholinguistics experiments (Alt, 2008). The similarities among the representations of colors and the corresponding ranked emotions in the latent similarity space have been compared on an English corpus. Table 1.1 lists the various colors and the ranked emotions they represent. To rank the emotions a 1-5 scale has been used.

Color	Ranking Emotions using Similarity with Color				
	Anger	Aversion/Disgust	Fear	Joy	Sadness
Blue	4	2	3	1	5
Red	4	3	2	1	5
Green	4	2	3	1	5
Orange	4	2	3	1	5
Purple	5	2	3	1	4
Yellow	4	2	3	1	5

Table 1.1: Ranked Emotions by Similarity with Colors (Strapparava and Ozbal, 2010)

In an experiment during the present work (Das and Bandyopadhyay, 2011; Das, 2011), it have been shown that sentimentality regarding any color changes with geo-spatial location and obviously with culture. The word "**blue**" evokes different sentiments for various cultures worldwide. The graphical illustration in Figure 1.1 may explain the situation better. The observation is that most of the negative sentiment regarding the word "**blue**" is from the middle-east and especially from the Islamic countries.

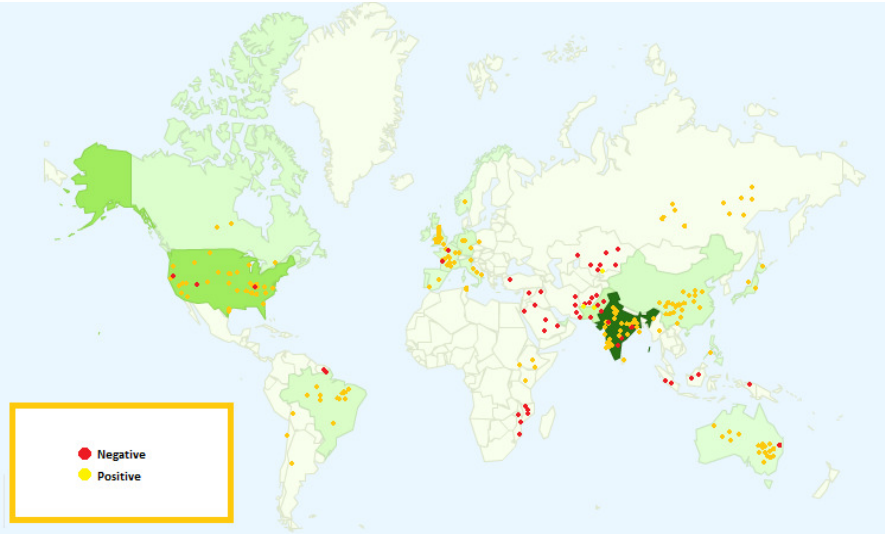


Figure 1.1: Geospatial Senti-Mentality

The following line from Wiki[2] (see in Religion Section) may provide a good explanation: "***Blue*** *in Islam: In verse 20:102 of the Qur'an, the word* زرق *zurq (plural of azraq 'blue') is used metaphorically for evil doers whose eyes are glazed with fear*". But there may be other possible explanations for this situation. But above all it is eminent that role of colors and culture changes with sentiment and sentiment also changes with color.

We have mentioned a few examples to show the difficulties regarding sentiment knowledge representation. The main issue is the various conceptual rules that govern sentiment and there are even more clues (possibly unlimited) that can map these concepts from realization to verbalization for a human being. For these reasons sentiment analysis is known as a multi-faceted (Liu, 2010) problem.

Previous studies have proposed multiple methods for creation of sentiment lexicons for multiple domains and languages. All these methods are based on storing prior polarity information with each lexical entry. The concept of prior polarity has been elaborated along with previous studies. The research endeavors in the present work follow next.

1.1.2 Prior Polarity: The Proposed Concept

In the previous sections the difficulties in lexicon level polarity assignment have been discussed. The current trend is to attach **prior polarity** to each entry in the sentiment lexicon. Prior polarity is an approximation based on statistics collected from corpus.

Prior polarity may be defined as the ratio between the distribution of the word as positive or negative in a corpus and the total occurrence of the particular word in the same corpus. Let us consider that the total occurrence of a word "**long**" in a domain corpus be *n*. The positive and negative occurrences of the word in the same corpus are S_p and S_n respectively.

Therefore in the developed sentiment lexicon the assigned positivity and negativity scores of the word will be as follows:

Positivity: $\frac{S_p}{n}$

Negativity: $\frac{S_n}{n}$

These prior polarity positivity and negativity values are approximate values. Prior polarity sentiment lexicons are necessary for a language to initiate the exploration of computational sentiment analysis for the language. Contextual polarity disambiguation techniques are still required for further sentiment/opinion analysis task.

[2] http://en.wikipedia.org/wiki/Blue

1.2 Related Works on Sentiment Knowledge Acquisition

In this section a detailed description of the previous studies is provided regarding sentiment knowledge acquisition for various domains and languages.

The development of the General Inquirer System[3] (1966) (Stone, 1966) by Philip Stones in Harvard was probably the first milestone to identify textual sentiment. They called it a content analysis research problem in the behavioral science. The aim was to gain understanding of the psychological forces and perceived demands of the situation that were in effect when the document was written. The system usually counts the positive or negative emotion instances. General Inquirer works with a stored list of manually-classified terms labeled with various positive or negative semantic orientations, and the words in the input are checked for their agreement or disagreement with the list. The lexicon used in the Inquirer system has been classified into several categories such as positive, negative, pleasure, feel, need, goal, place, vehicle etc. The rich lexical resource of General Inquirer[4] has been further used by various researchers to develop their sentiment or affect lexicon.

After the General Inquirer the community took a long break to start the current trend of sentiment analysis research. Jaynce Wiebe, Peter Turney and Vasileios Hatzivassiloglou are the pioneers who started the initial experimentations during early 90's. Jaynce Wiebe in 1990 (Wiebe, 1990) defines the term "**Subjectivity**" for Information Retrieval research but the term is now widely used by the sentiment research community. Any piece of text with sentiment relatedness is known as **subjective** whereas any piece of text that has no sentiment but factual statements is called **objective**. Later on in the year of 1997 Hatzivassiloglou (Hatzivassiloglou et. al., 1997) identified the semantic orientation of adjectives. This is the first research attempt that proves the effectiveness of empirical methods of building sentiment lexicon. After a few years Peter Turney (Turney, 2002) came up with his revolutionary approach of **Thumbs Up** and **Thumbs Down** for positive and negative review classification. Problem definition and formalization are one of the crucial steps of any scientific research. Researchers took nearly one and a half decade to formalize the sentiment analysis problem since early 90's to the first decade of this century. These research activities highlight the necessity of an automated system, which can identify the sentiment from electronic text documents.

(Hatzivassiloglou et. al., 1997) proposed the log-linear regression model to predict the orientation of conjoined adjectives. The log-linear regression model uses the number of constraints identified from a large corpus and clusters the conjoined adjectives into finite number of groups of different orientations. Finally, adjectives are labeled as positive or negative. The approach relies on some linguistic features, or indicators, with semantic orientation of conjoined adjectives that syntactically co-occur. The hypothesis is that the conjoined adjectives usually are of the same orientation, for example, ***fair*** and ***legitimate***, ***corrupt*** and ***brutal***. The system is trained on a large corpus to identify these relations to predict the semantic orientation of conjoined adjectives that are linguistically anomalous, i.e., where there is no

[3] http://www.wjh.harvard.edu/~inquirer/
[4] http://www.wjh.harvard.edu/~inquirer/homecat.htm

linguistic clue to identify the semantic orientation of any word. The situation is reversed for "***but***", which usually connects two adjectives of different orientations, for example, ***short*** but ***good***, ***far*** but ***comfortable***. The system identifies and uses this indirect information in the following stages:

1. All conjunctions of adjectives are extracted from the corpus along with relevant morphological relations.
2. A log-linear regression model combines information from different conjunctions to determine if each two conjoined adjectives are of same or different orientation. The result is a graph with hypothesized same- or different-orientation links between adjectives.
3. A clustering algorithm separates the adjectives into two subsets of different orientation. It places as many words of same orientation as possible into the same subset.
4. The average frequencies in each group are compared and the group with the higher frequency is labeled as positive.

The performance of the reported system is quite high. The important contributions of this milestone work are summarized below:

- The requirement of an automatic system for detecting the non-linguistic characteristics like semantic orientation of text is established.
- The task was based on the hypothesis that syntactically co-occurred adjectives demonstrate same semantic orientation although there are some exceptional cases for "but" and others.

(Turney, 2002) devised an algorithm to extract Pointwise Mutual Information (PMI) for consecutive words and their semantic orientation. The experimentation has been done on movie review corpus and thus the semantic orientations are referred to as "***thumbs up***" or "***thumbs down***" instead of positive or negative as in (Hatzivassiloglou et. al., 1997). The simple syntactic patterns for which PMI scores are calculated are described in the Table 1.2.

First Word	Second Word	Third Word(Not Extracted)
JJ	NN or NNS	Anything
RB, RBR, or RBS	JJ	not NN nor NNS
JJ	JJ	not NN nor NNS
NN or NNS	JJ	not NN nor NNS
RB, RBR, or RBS	VB, VBD,VBN, or VBG	Anything

Table 1.2: Syntactic Patterns of POS tags for Pointwise Mutual Information (PMI) Calculation (Peter Turney, 2002)

The POS tagger Brill Tagger (Brill, 1994)[5] has been used for the task. Phrases with adjective, adverb, noun and verbs depict semantic orientation. After these phrases have been extracted PMI algorithm is applied to determine their semantic orientation.

(Strapparava and Valitutti, 2004) started with WordNet domain synsets and semi-automatically expanded the list using various methods. The resource developed is called WORDNET-AFFECT[6], which is a linguistic resource for a lexical representation of affective knowledge. WORDNET-AFFECT was developed in two stages. The first step identifies the first "core" of affective synsets from English WordNet and the second step extends the core with the relations defined in WORDNET.

A method similar to (Turney, 2002) has been proposed by (Gamon et. al., 2005). They hypothesized that words with same semantic orientation co-occur whereas words with opposite semantic orientation does not co-occur at sentence level. They started with very small number of seed words and iterated multiple times with a Machine Learning based classifier and finally developed a good coverage sentiment lexicon for the particular domain.

(Read, 2005) has introduced three different problems for sentiment classification: Topic, Domain and Time dependency of sentiment polarity. In the present thesis, it has been shown that associative polarity score at lexicon level also changes with time.

(Taboda et. al., 2006) have successfully created a set of 1,719 adjectives whose Semantic Orientation (SO) was calculated using different methods. The authors proposed to extract the SO of 400 reviews from the website Epinions.com (about movies, music, books, hotels, cars, phones, computers, and cookware), using a weighted average of the adjectives in the texts. The final adjective dictionary contains 1,719 adjectives, whose SO was calculated using different methods: Altavista's NEAR (when it was available), Google's AND and extracting a subset of the positive/negative values from the General Inquirer (a total of 521 adjectives). The resource developed has been made free for further research through SentimentAI[7] group.

(Esuli and Sebastiani, 2006) have introduced the idea of SentiWordNet[8] that has established itself as the most widely used lexicon resources for sentiment analysis in the successive years. It is a semi-automatically developed lexical resource, which holds WordNet synsets and prior polarity scores as positivity and negativity scores. If the total occurrences of a word in a domain corpus is *n* and the positive and negative occurrences of that word are S_p and S_n respectively, then in a developed sentiment lexicon the assigned positivity and negativity scores of that word are defined following the as

[5] http://www.ling.gu.se/~lager/mogul/brill-tagger/index.html
[6] http://wndomains.fbk.eu/wnaffect.html
[7] http://groups.yahoo.com/group/SentimentAI/
[8] http://sentiwordnet.isti.cnr.it/

equation 1.1. Four years later in 2010, the authors have released the next version of the resource called SentiWordNet 3.0[9].

A nice architecture for the development of subjectivity lexicon from English to Romanian, a resource scarce language, has been proposed by (Wiebe and Mihalcea, 2007). The authors started with a small set of seed words for four POS categories - noun, verb, adverb and adjective. The list is incremented through bootstrapping using online dictionary and a small set of manually annotated corpora. The Subjectivity lexicon for English is one of the widely used English sentiment lexicon mainly developed from news corpora. The authors showed that lexico-syntactic patterns such as: X-Drive and Y-got-Angry help to identify subjective expressions across domains. A subjectivity classifier has been trained on a manually annotated data set and has been used on a truly unannotated data. The unannotated data have been used for training purposes by bootstrapping.

(Pang et. al., 2002) has suggested the building of sentiment lexicon manually for a domain. They involved two annotators independently to choose good indicator words for positive and negative sentiments in movie reviews. The responses are converted into simple decision procedures that essentially count the number of the proposed positive and negative words in a given document. These clue words are used as a seed for learning of other machine learning approaches like Naïve Bayes, Maximum Entropy and Support Vector Machine.

(Denecke, 2009) reported an interesting study on multiple domains to demonstrate the usefulness of the prior polarity scores from the SentiWordNet. The author proposed one rule-based and another machine learning based method. The positivity, negativity and the objectivity scores have been used from the SentiWordNet. The rule-based method achieved an accuracy rate of 74% which improved to 82% in the machine learning based approach. The author finally concluded that to use SentiWordNet scores effectively at sentence or phrase level they may need more sophisticated NLP techniques for good result.

(Mohammad et. al., 2009) proposed an automatic technique to increase the coverage of sentiment lexicon. The reported evaluation results show that the generated lexicon has high-coverage compared to SentiWordNet. The proposed technique captures both the individual words and multi-word expressions, using only a Roget-like thesaurus and a list of affixes. The authors have proposed two automatic methods, one is the automatic generation of antonymy and the second one is the Thesaurus based approach. A few hand-crafted rules have been proposed to generate more and more antonymy pairs. Table 1.3 lists some of these rules as patterns, the number of word-pairs generated and an example word pair. The thesaurus based technique examines each thesaurus paragraph for seed words. The seed word list is basically a manually augmented list. If a thesaurus paragraph has more positive seed words than negative seed words, then all the words (and multiword expressions) in that paragraph are marked as positive. Otherwise, all words in the paragraph are marked as negative.

[9] http://sentiwordnet.isti.cnr.it/

Affix	Pattern	word pairs	example word pair
X	disX	382	honest–dishonest
X	imX	196	possible–impossible
X	inX	691	consistent–inconsistent
X	malX	28	adroit–maladroit
X	misX	146	fortune–misfortune
X	nonX	73	sense–nonsense
X	unX	844	happy–unhappy
X	Xless	208	gut–gutless
lX	illX	25	legal–illegal
rX	irX	48	responsible–irresponsible
Xless	Xful	51	harmless–harmful
Total		2692	

Table 1.3: Orthographic Antonymy Generation Rules (Mohammad, et al., 2009)

(Mohammad and Turney, 2010) suggested Amazon Mechanical Turk, an online service from Amazon, to obtain a large amount of human annotation of emotion lexicon in an efficient and inexpensive manner. However, the task must be carefully defined to obtain high quality annotations. Several checks are necessary to ensure that random and erroneous annotations are discouraged, rejected, and re-annotated. By this process only 2081 words are tagged with an average tagging of about 4.75 tags per word.

1.3 Sentiment Lexicon Acquisition: the Work Done

The present work started with the sentiment lexicon acquisition for Bengali language. This is the first work on sentiment analysis in Bengali. Several experimentations have been carried out for sentiment lexicon generation for Bengali. Various automatic processes like bilingual dictionary based, WordNet based synonym and antonym expansion, orthographic antonym generation and corpus based approach have been explored to generate the lexical resource from a resource rich language like English.

There are two qualitative aspects of sentiment lexicon acquisition: **Coverage** and **Credibility**. The reported techniques for creation of Sentiment Lexicon in several languages and domains can be broadly categorized into two classes, one follows the classical manual annotation techniques (Andreevskaia and Bergler, 2006; Wiebe and Riloff, 2005; Mohammad et. al., 2008) while the other follows various automatic techniques (Tong, 2001). Both types of techniques have few limitations. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain and language. Manually augmented resources are undoubtedly trustable but it generally takes time to build. Manual annotation techniques require a large number of annotators to balance one's sentimentality in order to reach agreement. But human annotators are quite unavailable and costly.

It has been observed during the present work that automatic processes are trustable for sentiment lexicon **generation** or **coverage expansion** but still manual methods are needed as sentiment is a property of human intelligence and is not entirely based on the features of a language. Thus human involvement is necessary to capture the sentiment of the society. But there is scarcity of human annotators and the manual method takes time and cost. Therefore an online game has been developed to attract internet population for automatic collection of sentiment polarity knowledge. The developed online game "***Dr. Sentiment***", revolutionize the idea of creating prior polarity sentiment lexicon for any new language (presently 57) by involving internet population. This technique also helps to attract human annotators (players) with literally zero cost! It has been established in the present work that the proposed methods may be replicated for other languages as well.

1.3.1 Source Language Lexicon Acquisition

Several prior polarity sentiment lexicons are available for English: SentiWordNet (Esuli et. al., 2006), Subjectivity Word List (Wilson et. al., 2005), WordNet Affect list (Strapparava et al., 2004) and Taboada's adjective list (Taboada et al., 2006). SentiWordNet and Subjectivity Word List have been identified as the most reliable source lexicons. The first one is widely used and the second one is robust in terms of performance. Taboada's adjective list is not considered in the present work as it contains only 1,719 adjectives. WordNet Affect list is also not considered as it contains emotion information. Various statistics of the English SentiWordNet and Subjectivity Word List are reported in Table 1.4.

Words	**SentiWordNet**		**Subjectivity Word List**	
Total Entries	**Single**	**Multi**	**Single**	**Multi**
	115424	79091	5866	990
Unambiguous	20789	30000	4745	963
Ambiguous	**Threshold < 0.4**		**Subjectivity Strength (low)**	**POS (anypos)**
	86944	30000	2652	928

Table 1.4: A Closer Look on SentiWordNet and Subjectivity Word List

A merged English sentiment lexicon has been generated from both the SentiWordNet and the Subjectivity Word List after removing the duplicates and applying other filtering techniques. It has been observed that **64%** of the single word entries are common in both the existing resources. The generated sentiment lexicon contains **14,135** numbers of tokens.

A subset of 8,427 sentiment words has been extracted from the English SentiWordNet, by selecting those whose orientation strength is above the heuristically identified threshold of **0.4**. The words whose orientation strength is below 0.4 are ambiguous and may lose their subjectivity in the target language after translation. A total of 2652 weakly subjective words are discarded from the Subjectivity word list. For this task, the same technique as proposed by (Rada et al., 2007) has been followed.

In the next stage the words whose POS categories in the Subjectivity word list are undefined and have been tagged as "**anypos**" are discarded. These words may generate sense ambiguity issues in the next stages of sentiment analysis.

Some words in the Subjectivity word list are inflected, e.g., ***memories***. These words would be stemmed during the translation process, but some words present no subjectivity property after stemming (***memory*** has no subjectivity property). A word may occur in the subjectivity list in many inflected form like ***zeal, zealot, zealous, zealously***. Individual clusters for the words sharing the same root form are created and the root form is further checked in the SentiWordNet for validation. If the root word exists in the SentiWordNet then it is assumed that the word remains subjective after stemming and hence is added to the new list. Otherwise, the cluster is completely discarded to avoid any further ambiguities. Details could be found in (Das and Bandyopadhyay, 2010(c)) and (Das and Bandyopadhyay, 2010(e)).

1.3.2 Automatic Generation and Expansion of Sentiment Lexicon

Four types of automatic processes have been proposed for generation and expansion of sentiment lexicon for three Indian languages (Hindi, Bengali and Telugu). The automatic processes are Dictionary based, WordNet Based Synonym Expansion & Antonym Expansion and Antonym Generation and finally monolingual Corpus based expansion technique. The automatic processes mainly contribute towards increasing the coverage of sentiment lexicons. These automatic processes generate lexicons for target languages and the prior polarity scores are copied from the source English sentiment lexicon. The contribution of each automatic process in coverage expansion differs from language to language. The reasons appear to be the linguistics resources used for the language, like, dictionary, WordNet and corpus. The automatic processes have been designed mainly for Bengali and later on the processes (Das and Bandyopadhyay, 2010(c); Das and Bandyopadhyay, 2010(e)) have been adopted for two other Indian languages, Hindi and Telugu. Hindi is the national language of India and it is the fourth language in the world in terms of the number of native speakers. Telugu is a south Indian language and the total number of Telugu speaker is approximately 75 million[10]. It is hoped that these techniques can be adopted for any new languages as well. Bengali SentiWordNet[11] generated by these processes is already made available to the academic and research community for research purposes only.

1.3.2.1 Dictionary Based Approach

A word-level translation process followed by error reduction technique has been adopted for generating the Indian languages SentiWordNet(s) from the English sentiment lexicon which is developed by merging the English SentiWordNet and the Subjectivity Word List.

English to Indian language synsets are being developed under the national level Project "Development of English to Indian Languages Machine Translation Systems (EILMT)[12]", a consortia project funded by

[10] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
[11] http://www.amitavadas.com/sentiwordnet.php
[12] http://www.cdacmumbai.in/e-ilmt

Department of Information Technology (DIT), Government of India. These synsets are robust and reliable as these are created by native speakers as well as linguistics experts of the specific languages. The first phase of the project has been executed for the following Indian languages (Hindi, Bengali, and Telugu). For each language, approximately 9966 synsets are available along with the English WordNet offset. These bilingual synset dictionaries have been used along with language specific dictionaries.

A word level synset/lexical transfer technique is applied to each English synset/word in the merged sentiment lexicon. Each dictionary search produces a set of Indian language synsets/words for a particular English synset/word.

1.3.2.1.1 Hindi

Two available manually compiled English-Hindi electronic dictionaries have been identified for the present task. The first dictionary is the SHABDKOSH[13] and the second one is the Shabdanjali[14]. These two dictionaries have been merged automatically by removing the duplicates. The merged English-Hindi dictionary contains approximately 90,872 unique entries. The new merged resource has been used for English to Hindi sentiment lexicon generation via cross lingual projection. The prior polarity sentiment scores for the Hindi synset/words are copied from the source English sentiment lexicon.

The bilingual dictionary based translation process has resulted in 22,708 Hindi entries in the Hindi sentiment lexicon.

1.3.2.1.2 Bengali

An English-Bengali dictionary (approximately 102119 entries) has been developed using the Samsad Bengali-English dictionary[15]. The English-Bengali bilingual dictionary has been successfully used for the translation of the English sentiment lexicon entries. The prior polarity sentiment scores for the Bengali synset/words are copied from the source English sentiment lexicon equivalents.

The bilingual dictionary based translation process has resulted in 35,805 Bengali entries. A manual checking is done to identify the reliability of the words generated through this automatic process. After manual checking only 1688 words are discarded, i.e., the final list consists of 34,117 words.

1.3.2.1.3 Telugu

Charles Philip Brown English-Telugu Dictionary[16], the Aksharamala[17] English-Telugu Dictionary and the English-Telugu Dictionary[18] developed by Language Technology Research Center (LTRC), International Institute of Hyderabad (IITH) have been chosen for the present task. There is no WordNet publicly

[13] http://www.shabdkosh.com/
[14] http://www.shabdkosh.com/content/category/downloads/
[15] http://dsal.uchicago.edu/dictionaries/biswas_bengali/
[16] http://dsal.uchicago.edu/dictionaries/brown/
[17] https://groups.google.com/group/aksharamala
[18] http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html

available for Telugu and the Telugu corpus used is very small in size. Therefore the dictionary based approach is the main process for Telugu SentiWordNet generation.

These three dictionaries have been merged automatically after removing the duplicates. The merged English-Telugu dictionary contains approximately 112310 unique entries. The English-Telugu bilingual dictionary has been successfully used for the translation of the English sentiment lexicon entries. The prior polarity sentiment scores for the Telugu words are copied from their English sentiment lexicon equivalents.

The dictionary based translation process has resulted in 30,889 Telugu entries, about 88% of final Telugu SentiWordNet synsets.

1.3.2.2 WordNet Based Approach

WordNet is treated as the best monolingual lexical resource in NLP activities. The WordNet has been chosen for expanding the lexicons generated by the bilingual dictionary based process. Synonym and antonym based expansion techniques have been developed for both the source and target languages.

English synsets/words from merged source lexicon are checked into the English WordNet[19] for equivalent synonyms and antonyms. The extracted synonyms and antonyms are then manually checked to confirm that they have sentiment orientation. Finally these synsets/words are translated into the target language by using bilingual dictionaries as described in the previous section. This process expands approximately 12% of the source language lexicon and near about 8% for Bengali and Hindi and 9% of Telugu lexicons after the bilingual dictionary based translation.

WordNet based expansion technique has been developed for Bengali and Hindi target languages only. No Telugu WordNet is publicly available.

Prior polarity scores for expanded synonyms are kept the same for all the members of the synset. The calculated prior polarity positivity and negativity scores for antonym synsets are calculated as:

$$T_p = 1 - S_p$$
$$T_n = 1 - S_n \quad \text{----- (1.1)}$$

where S_p, S_n are the positivity and negativity scores for the source synsets (i.e., English) and T_p, T_n are the positivity and negativity scores for target synsets for any language (i.e., Hindi and Bengali) respectively.

[19] http://wordnet.princeton.edu/

1.3.2.2.1 Hindi

Hindi WordNet[20] (Jha et al., 2001) is a well structured and manually compiled resource and is continuously being updated since the last nine years. There is an available API[21] for accessing the Hindi WordNet. Almost 60% of final SentiWordNet synsets in Hindi are generated by this method.

1.3.2.2.2 Bengali

The Bengali WordNet[22] (Robkop et al., 2010) is being developed by the Asian WordNet (AWN) community. It contains 1775 noun synsets only as reported in (Robkop et al., 2010). A Web Service[23] has been provided for accessing the Bengali WordNet. There are only a few number of noun synsets in the Bengali WordNet. Other important POS category words for sentiment lexicon such as adjective, adverb and verb are absent. Only 5% new Bengali SentiWordNet lexicon entries have been generated through this process.

1.3.2.3 Antonymy Expansion

Automatically or manually created lexicons have limited coverage and do not include most semantically contrasting word pairs (Mohammad et. al., 2009).

Affix/Suffix	Word	Antonym
***ab*X**	Normal	*Ab*-normal
***mis*X**	Fortune	*Mis*-fortune
***im*X-*ex*X**	*Im*-plicit	*Ex*-plicit
***anti*X**	Clockwise	*Anti*-clockwise
***non*X**	Aligned	*Non*-aligned
***in*X-*ex*X**	*In*-trovert	*Ex*-trovert
***dis*X**	Interest	*Dis*-interest
***un*X**	Biased	*Un*-biased
***up*X-*down*X**	*Up*-hill	*Down*-hill
***im*X**	Possible	*Im*-possible
***ill*X**	Legal	*Il*-legal
***over*X-*under*X**	Overdone	*Under*-done
***in*X**	Consistent	*In*-consistent
***r*X-*ir*X**	Regular	*Ir*-regular
X*less*-X*ful*	Harm-*less*	Harm-*ful*
***mal*X**	Function	*Mal*-function

Table 1.5: Rules for Generating Orthographic Antonyms

To overcome this limitation and to increase the coverage of the SentiWordNet(s) an automatic antonymy generation technique is applied followed by corpus validation to check whether the orthographically generated antonym does really exist. Only 16 hand crafted rules have been used as reported in Table 1.5. These rules are first used to expand the original English lexicon and to translate by

[20] http://www.cfilt.iitb.ac.in/wordnet/webhwn/
[21] http://www.cfilt.iitb.ac.in/wordnet/webhwn/API_downloaderInfo.php
[22] http://bn.asianwordnet.org/
[23] http://bn.asianwordnet.org/services

dictionary look up for corresponding languages, i.e., Hindi, Bengali and Telugu. About 8% of Bengali, 7% of Hindi and 11% of Telugu SentiWordNet entries are generated by this process.

1.3.2.4 Corpus Based Approach

Language/culture specific words are to be captured in the SentiWordNet for good coverage. The sentiment lexicon generation techniques via cross-lingual projection are unable to capture these words. For example, the following words are language specific sentiment words:

सहेरा (Sahera: A marriage-wear)

দুর্গাপুজো (Durgapujo: A festival of Bengal)

To increase the coverage of the developed SentiWordNet(s) and to capture the language/culture specific words an automatic corpus based approach has been proposed. At this stage the developed SentiWordNet(s) for the three Indian languages have been used as a seed list. The language specific corpus is automatically tagged with SWP (Sentiment Word Positive) and SWN (Sentiment Word Negative) tags for the seed words. Although the words in the seed list have both positivity and negativity scores r a word level tag is preferred as either positive or negative based on the highest sentiment score.

A Conditional Random Field (CRF[24]) based Machine Learning model is then trained with the seed list corpus along with multiple linguistics features such as morpheme, parts-of-speech and chunk label. These linguistics features have been extracted by the shallow parsers[25] for Indian languages. An n-gram (*n*=4) sequence labeling model has been used for the present task.

The monolingual corpora used have been developed under Project "Development of English to Indian Languages Machine Translation Systems (EILMT) Systems". Each corpus has approximately 10K of sentences.

1.3.3 Involving Human Intelligence

There are several motivations behind the development of the intuitive game to automatically collect human sentiment oriented information at the lexicon level. In the history of Information Retrieval research there is a milestone when ESP game[26] (Ahn et al., 2004) innovate the concept of a game to automatically label images available in the World Wide Web. It has been identified as the most reliable strategy to automatically annotate the online images. The success of the Image Labeler game has motivated the present work.

A number of research endeavors could be found in the literature for creation of Sentiment Lexicon in several languages and domains. These techniques can be broadly categorized into two genres; one

[24] http://crfpp.sourceforge.net
[25] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php
[26] http://www.espgame.org/

follows the classical manual annotation techniques while the other follows various automatic techniques. Both types of techniques have few limitations. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain. Manually augmented resources are undoubtedly trustable but it generally takes time to build. Manual annotation techniques require a large number of annotators to balance the sentimentality of an individual to reach agreement. But human annotators are not easily available and are quite costly.

Sentiment lexicon generation with cross lingual projection has several issues or limitations, i.e.,

- Source language word may have no sentiment value in the target language
- Sentiment score in the target language may not be the same as in the corresponding source language word
- Language / Culture specific entries should be included in the lexicon
- Sentiment score should be dynamic in the time scale, i.e., it should be updated with time.
- **Relative sentiment score** is needed rather than fixed point probabilistic score.

Now what are relative sentiment scores? For example, both the words ***good*** and ***better*** are positive but we need to know which one is more positive than the other. Table 1.6 shows how in SentiWordNet relative scoring has been made. This relative scoring is necessary for more fine grained sentiment detection.

Word	Positivity	Negativity
Good	0.625	0.0
Better	0.875	0.0
Best	0.980	0.0

Table 1.6: Relative Sentiment Scores in SentiWordNet

It has been observed that the automatic techniques for sentiment lexicon generation mainly help in increasing the coverage of the target language lexicon. Sentiment reflects human intelligence and is not entirely based on the features of a language. Human involvement is necessary to capture the sentiment orientation of the society that may be mapped to the prior polarity score.

With this hypothesis an online game has been developed to attract the internet population for automatic accumulation of the sentiment lexicon. Involvement of Internet population for lexical resource generation is an effective approach as the population is very high in number and ever growing

(approx. 360,985,492)[27]. The NLP community always faces the bottleneck of classical human annotation due to unavailability of human annotators, slow annotation speed and investment of large money. But human annotation is trustable and thus unavoidable. Therefore the target was to minimize the hidden costs: unavailability, speed, money etc. Another important problem that surfaces during sentiment annotation is that the human annotators may be sentimentally biased. For example, "***eating_octopus***" may be positive for Thai people but it is not the same for rest of the world. Therefore the socio, economic, religious, and cultural and many more unknown hidden aspects control the sentiment of a human being. This situation demands that the sentiment annotators should be diverse in nature. Internet population is the best choice in this respect as it consists of people with various languages, cultures, age etc and thus is not biased towards any domain, language or particular society. Table 1.7 confirms the diverse nature of the Internet population.

WORLD INTERNET USAGE AND POPULATION STATISTICS						
World Regions	**Population (2010 Est.)**	**Internet Users Dec. 31, 2000**	**Internet Users Latest Data**	**Penetration (Population)**	**Growth 2000-2010**	**Users % of Table**
Africa	1,013,779,050	4,514,400	110,931,700	10.9 %	2,357.3 %	5.6 %
Asia	3,834,792,852	114,304,000	825,094,396	21.5 %	621.8 %	42.0 %
Europe	813,319,511	105,096,093	475,069,448	58.4 %	352.0 %	24.2 %
Middle East	212,336,924	3,284,800	63,240,946	29.8 %	1,825.3 %	3.2 %
North America	344,124,450	108,096,800	266,224,500	77.4 %	146.3 %	13.5 %
Latin America/Caribbean	592,556,972	18,068,919	204,689,836	34.5 %	1,032.8 %	10.4 %
Oceania / Australia	34,700,201	7,620,480	21,263,990	61.3 %	179.0 %	1.1 %
WORLD TOTAL	6,845,609,960	360,985,492	1,966,514,816	28.7 %	444.8 %	100.0 %

Table 1.7: Internet Usage and Population Statistics

The sentiment lexicon generated through the internet population is credible as it is tagged by human annotators. It is not a static sentiment lexicon as it is updated regularly. Around 10-20 players each day are playing it throughout the world in different languages. The average number of tagging per word is about 7.47 till date.

[27] http://www.internetworldstats.com/stats.htm

1.3.3.1 Dr. Sentiment

Dr. Sentiment[28] (Das and Bandyopadhyay, 2011);(Das , 2011) is a template based interactive online game, which collects player's sentiment by asking a set of simple template based questions and finally reveals a player's sentimental status. Dr. Sentiment fetches random synset/word from the merged source lexicon and asks every player to tell about his/her sentiment polarity understanding of the concept behind the word fetched by it.

The gaming interface has four types of question templates. The question templates are named as Q1, Q2, Q3 and Q4. To make the gaming interface more interesting images have been added. These images have been retrieved by Google image search API[29]. But images could trigger biased sentiment in the players. Figure 1.2 shows an image retrieved by Google with the input word "**Heavy**". Overall it presents an ambiguity for the sense of the word "**Heavy**" whereas the word generally triggers a sense of positivity in our mind. The image generates a sense of negativity whereas the original word is positive (*Heavy Metal*)! The corresponding polarity score of the word "**Heavy**" from SentiWordNet (English) are reported in the following example. To avoid biasness randomization has been among the first ten images retrieved by Google.

Figure 1.2: Image of "***Heavy***": Misleading Sentiment

POS	OFFSET	Positivity	Negativity	Lexicon
Adjective	1102371	0.625	0.0	Heavy

[28] http://www.amitavadas.com/Sentiment%20Game/
[29] http://code.google.com/apis/imagesearch/

The Sign Up form of the "Dr. Sentiment" game asks the player to provide personal information such as Sex, Age, City, Country, Language and Profession. These collected personal details of a player are stored as a log record in the database.

Dr. Sentiment has proved itself as an excellent technique to collect the sentiment of the society. The gaming architecture is quite generic and thus a decision has been taken to make it multilingual. Dr. Sentiment presently supports 57 languages (with the help of Google Translation API[30]) including English. The SentiWordNet created by Dr. Sentiment is identified as the **Global SentiWordNet** (Das and Bandyopadhyay, 2010(d)). The languages presently supported by Dr. Sentiment are reported in the Table 1.8.

Languages							
Afrikaans	Bulgarian	Dutch	German	Irish	Malay	Russian	Thai
Albanian	Catalan	Estonian	Greek	Italian	Maltese	Serbian	Turkish
Arabic	Chinese	Filipino	Haitian	Japanese	Norwegian	Slovak	Ukrainian
Armenian	Croatian	Finnish	Hebrew	Korean	Persian	Slovenian	Urdu
Azerbaijani	Creole	French	Hungarian	Latvian	Polish	Spanish	Vietnamese
Basque	Czech	Galician	Icelandic	Lithuanian	Portuguese	Swahili	Welsh
Belarusian	Danish	Georgian	Indonesian	Macedonian	Romanian	Swedish	Yiddish

Table 1.8: The Languages Covered by the Global SentiWordNet

1.3.3.2 Strategy

Dr. Sentiment asks 30 questions to each player. There are predefined distributions of each question type as 11 for Q1, 11 for Q2, 4 for Q3 and 4 for Q4. But these predefined distributions and the total number of questions could be changed for more experimentations. The questions from each question type are randomly asked to keep the game more interesting. Additionally a log record has been kept with every player's session to ensure that no word is repeated. At each Question (Q) level translation service[31] has been used to display the sentiment word in the language of the player. Google translation service has been used for word based translation. Google API provides multiple word level translations corresponding to different senses but currently only the first sense is selected automatically.

This type of gaming technique to collect language resources becomes successful when more and more players play the game. The most important motivating factor to the players is that Dr. Sentiment can reveal their sentimental status: whether they are extreme negative or extreme positive or very much neutral or diplomatic etc. It is not claimed that the revealed sentiment status of a player by Dr. Sentiment is exact or ideal. It is only to make the players motivated but the outcomes of the game

[30] http://www.google.com/language_tools?hl=EN
[31] http://translate.google.com/

effectively helps to store human sentimental psychology in terms of computational lexicon. Randomizing the question selection and wide range of varieties in the comments keep the game interesting so that players occasionally return to play the game. Around 10-20 players are playing the game each day throughout the world in different languages. The average number of taggings per word is about 7.47 till date. In the following sub sections the strategy for each question type are discussed. Some snaps from Dr. Sentiment are shown in the Figure 1.3.

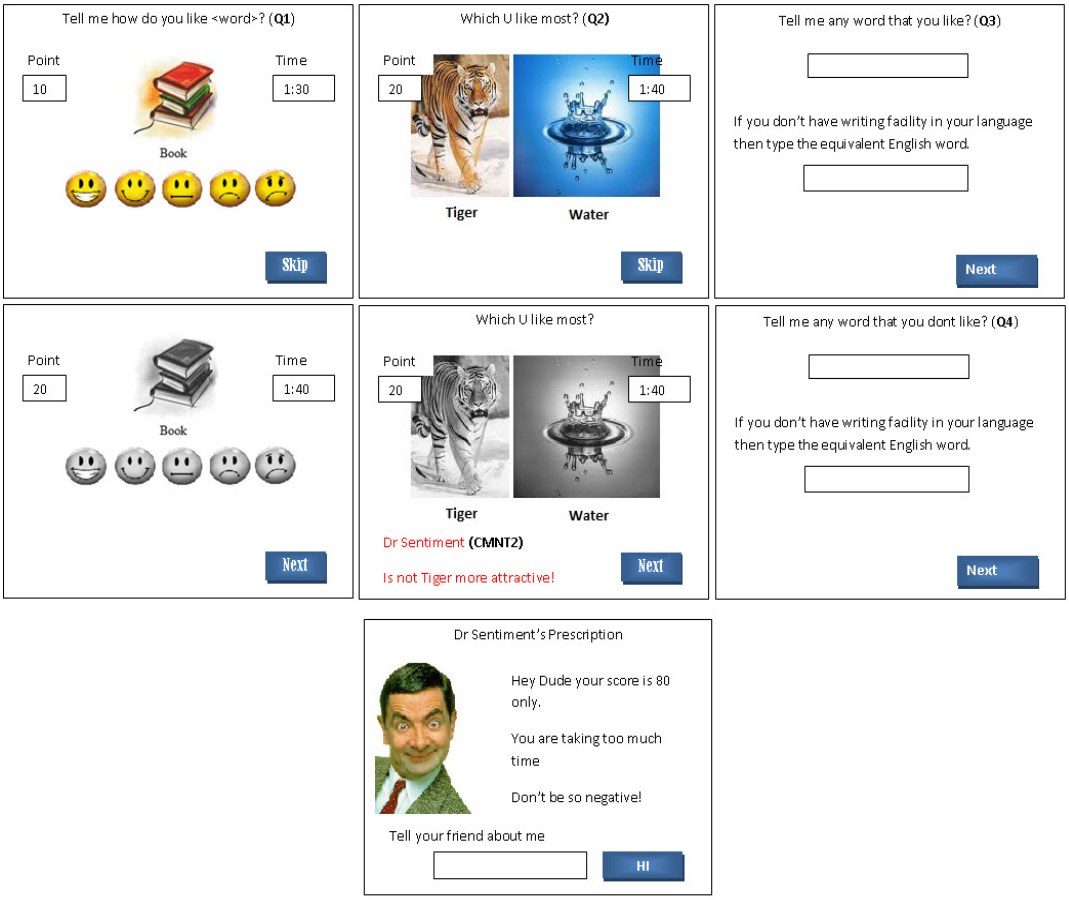


Figure 1.3: Snaps from the Dr. Sentiment Game

1.3.3.2.1 Q1

An English word from the English SentiWordNet is randomly chosen. A Google image search API is fired with the word as a query. An image along with the word itself is shown in the Q1 page of Dr. Sentiment game to make it more attractive. The words are shown in the language of the player as specified in the login page.

Players press the different emoticons (Figure 1.4) to express their sentimentality. The interface keeps log records of each clicking interaction. The sentiment scores are calculated by the different emoticons pressed by various players with the following scale of sentiment scores as extreme positive (pos: 0.5,

neg: 0.0), positive (pos: 0.25, neg: 0.0), neutral (pos: 0.0, neg: 0.0), negative (pos: 0.0, 0.25) and extreme negative (pos: 0.0, neg: 0.5).

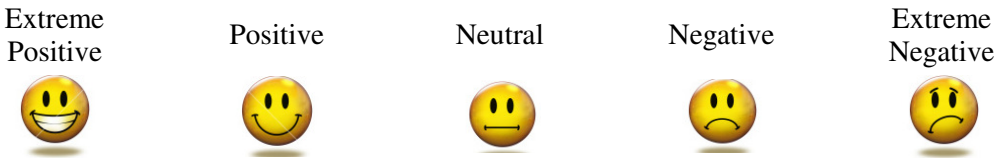


Figure 1.4: Emoticons as They Appear in Dr. Sentiment Game

For Languages other than English the word along with its associated properties (POS, Offset: The WordNet Offset) are inserted into the language table (i.e., SentiWordNet for the corresponding language). The new positivity and negativity scores are stored over the original scores on the English word and finally copied to the language table. These language tables are the corresponding SentiWordNet(s) for each language.

1.3.3.2.2 Q2

This question type is specially designed for identifying the relative score to a word. For example, both ***good*** and ***better*** are positive but still we need to know which one is more positive than the other. Table 1.6 shows how relative scoring has been made in SentiWordNet. With the present gaming technology relative polarity scoring has been assigned to each ***n-n*** word pair combination.

Randomly ***n*** (presently 2-4) words have been chosen from each source SentiWordNet synset along with their images as retrieved by Google API. Each player is then asked to select one of these selected words that he/she likes most. The relative score is calculated and stored in the corresponding log table. The mechanism of relative score calculation is very simple. Suppose we have two words w_1 and w_2. If a player chooses w_1 over w_2 then the relative polarity mechanism is as follows:

- If positivity score of w_1 is higher than w_2 or negativity score of w_2 is higher than w_1 then no update has been done.
- If positivity score of w_1 is lower than w_2 then $P_{w_1} += \frac{p_{w_2} - P_{w_1}}{2}$ ----- (1.2)
- If negativity score of w_2 is lower than w_1 then $N_{w_2} += \frac{N_{w_1} - N_{w_2}}{2}$ ----- (1.3)

It may be an interesting question to know what happens if the positivity scores of w_1 remains lower than w_2 after operation (1.2) or the negativity score of w_2 remains lower than w_1 after operation (1.3)? The relative scores are continuously updated by several players and it is a bad idea to rely on a single player, because he/she may be biased towards those concepts (i.e., w_1 and w_2). If other players choose w_1 over w_2 iteratively then after a few iterations automatically the positivity score of w_1 will be higher than w_2 or the negativity score of w_2 will be higher than w_1.

1.3.3.2.3 Q3

The player is asked for any positive word in his/her mind. The Q3 and Q4 type of questions help to increase the language/culture specific word coverage of the existing sentiment lexicon. The word is then added to the existing sentiment lexicon and further used for Q1 or Q2 type questions to know the sentimentality of other users about the particular word. The following example shows language / culture specific positive sentiment words in Hindi and Bengali respectively.

सहेरा$^{pos+}$ (Sahera: A marriage-wear)

দুর্গাপুজো$^{pos+}$ (Durgapujo: A festival of Bengal)

1.3.3.2.4 Q4

A player is asked any negative word. The rest of the technique is the same as that for Q3 type questions. The following example shows a language / culture specific negative sentiment word in Hindi.

बन्ध$^{neg-}$ (Political closing of market places and transportation)

1.3.3.3 Comment Architecture

Comments are made to make the game interactive or interesting. The dynamic comments make the player feel that as if Dr. Sentiment is interacting with them. There are three types of Comments, Comment type 1 (CMNT1), Comment type 2 (CMNT2) and the final comment as Dr. Sentiment's prescription. CMNT1 and CMNT2 type comments are associated with question types Q1 and Q2 respectively.

1.3.3.3.1 CMNT1

Comment type 1 has 5 categories as reported in Table 1.9. CMNT1 is used for Q1. The five categories are:

- Positive word has been tagged as negative (**PN**)
- Positive word has been tagged as positive (**PP**)
- Negative word has been tagged as positive (**NP**)
- Negative word has been tagged as negative (**NN**)
- Neutral (**NU**)

Comments are randomly retrieved from the comment type table according to their category.

1.3.3.3.2 CMNT2

The strategy for CMNT2 type comments is the same as that for comment type CMNT 1. CMNT2 is used for Q2. Comment type 2 has only 2 categories:

- Positive word may have been tagged as negative. (PN)
- Negative word may have been tagged as positive. (NP)

PN	PP	NP	NN	NU
You don't like <word>!	Good you have a good choice!	Is <word> good!	Yes <word> is too bad!	You should speak out frankly!
You should like <word>!	I love <word> too!	I hope it is a bad choice!	You are quite right!	You are too diplomatic!
But <word> is a good itself!	I support your view!	I don't agree with you!	I also don't like <word>!	Why you are hiding from me? I am Dr. Sentiment.

Table 1.9: Comment Architecture in Dr. Sentiment

1.3.3.3.3 Dr. Sentiment's Prescription

Dr. Sentiment's prescription is the revealed sentimental status of a player by Dr. Sentiment. The motivating message for players is that Dr. Sentiment can reveal their sentimental status: whether they are extreme negative or positive or very much neutral or diplomatic etc. The final prescription for a player depends on various factors such as the total number of positive, negative or neutral comments a player receives during the session and the total time taken by the player. The final prescription also depends on the range of the accumulated values (like 10-20, 20-30 or else) of all the above factors.

A Facebook version of the Dr. Sentiment[32] game has been developed where the players can play. The final result gets automatically posted on the wall message of the player. This has been done to promote the game and to increase the number of players playing the game. The general idea is that people will see his/her friend playing the game therefore they will also be interested to play the game.

A word previously tagged by a player is avoided by the tracking system during subsequent turns by the same player. The objective is to tag more and more words involving Internet population. It has been observed that the strategy helps to keep the game interesting as a large number of players return to play the game after this strategy was implemented.

[32] http://www.facebook.com/?ref=home#!/pages/Dr-Sentiment/148683515193432

1.4 Sentiment Knowledge: Unexplored Dimensions

Sentiment analysis is one of the most explored research areas since the last few decades. Although a formidable amount of research has been done, the reported solutions and available systems do not meet the satisfaction level of the end user. The main issue is the various conceptual rules that govern sentiment and there are even more clues (possibly unlimited) that can convert these concepts from realization to verbalization of a human being. Human psychology directly relates to the paradigms of social psychology, culture, pragmatics etc. and governs the sentiment realization of us. Proper incorporation of human psychology into computational sentiment knowledge representation appears to be the step in the right direction.

Dr. Sentiment helps to collect not only sentiment polarity at lexicon level but also social psychology with various aspects. It has been already mentioned that the summation of all the regulating aspects of sentiment orientation is human psychology and thus it is a multi-faceted problem (Liu, 2010). The governing aspects wrapped in the present sentiment lexicon are **Gender**, **Age**, **City**, **Country**, **Language** and **Profession**. The Sign Up form of the "Dr. Sentiment" game asks the player to provide personal information such as Sex, Age, City, Country, Language and Profession. This information is effectively kept as a log record with the lexicon. Additional psychological aspects could be added in future to hold the human psychology.

The developed SentiWordNet(s) wrapped with psychological information is called PsychoSentiWordNet. The **PsychoSentiWordNet** (Das, 2011) is an extension of the existing SentiWordNet 3.0 (Baccianella et. al., 2010) to hold the possible psychological aspects. The PsychoSentiWordNet holds variable prior polarity scores that could be fetched depending upon these psychological aspects. An example with the input word '**High**' may illustrate the definition better:

Aspects (Profession)	Polarity
Null	Positive
Businessman	Negative
Share Broker	Positive

The information collection can be possible for other dimensions as well. It is expected that the generated lexicon will analyze sentiments better over fixed point prior polarity lexicons. We find that this is the first endeavor where sentiment analysis meets Artificial Intelligence (AI) and psychology. The following sub-sections describe the detailed impact of every psychological aspect to understand the social psychology.

1.4.1 Senti-Mentality

Several analyses have been done on the developed sentiment lexicons to understand the sentimental behavior of people depending upon various psychological aspects. Statistical analyses reveal some interesting observations that support the effectiveness of the generated lexicons.

1.4.1.1 Geospatial Senti-Mentality

During analysis we had an interesting observation. The word "**blue**" gets tagged by different players around the world. But surprisingly it has been tagged as positive from one part of the world and negative from a different part of the world. The graphical illustration in Figure 1.5 illustrates the problem. Most of the negative taggings are coming from middle-east and especially from Islamic countries. While analyzing this peculiar behavior we found the following line in Wiki[33] (see in Religion Section): "Blue in Islam: In verse 20:102 of the Qur'an, the word زرق zurq (plural of azraq 'blue') is used metaphorically for evil doers whose eyes are glazed with fear".

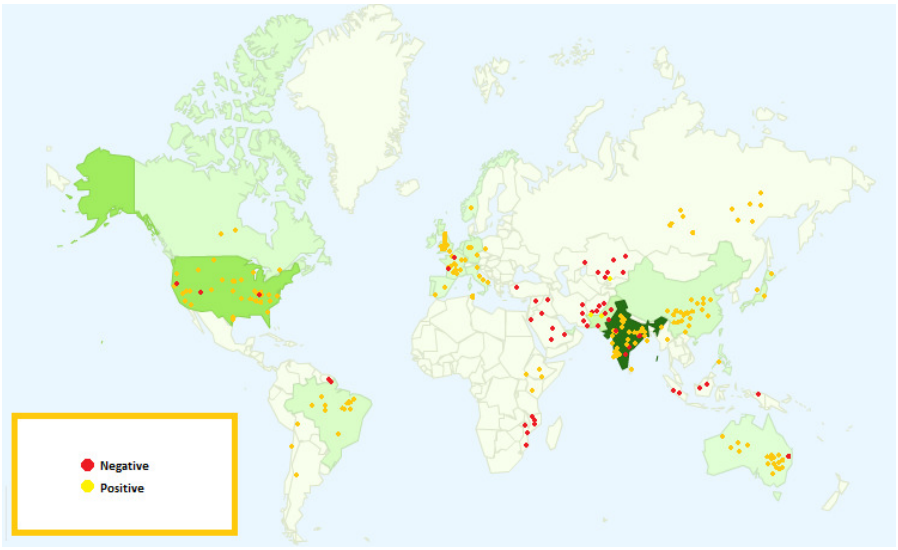


Figure 1.5: Geospatial Senti-Mentality

1.4.1.2 Age Wise

Another interesting observation is that sentiment understanding may vary age-wise. Total 533 players have played the game till date. The age wise distribution of the players is shown in the Figure 1.6. The number of players under each age group is shown at the top of every vertical bar. The vertical bars are divided into two colors (Green depicts Positivity and Red depicts negativity) that show the percentage of the players in that age group who have responded with positive and negative scores during playing. It gives an idea of the overall change of senti-mentality of a human being during various stages of his/her life.

[33] http://en.wikipedia.org/wiki/Blue

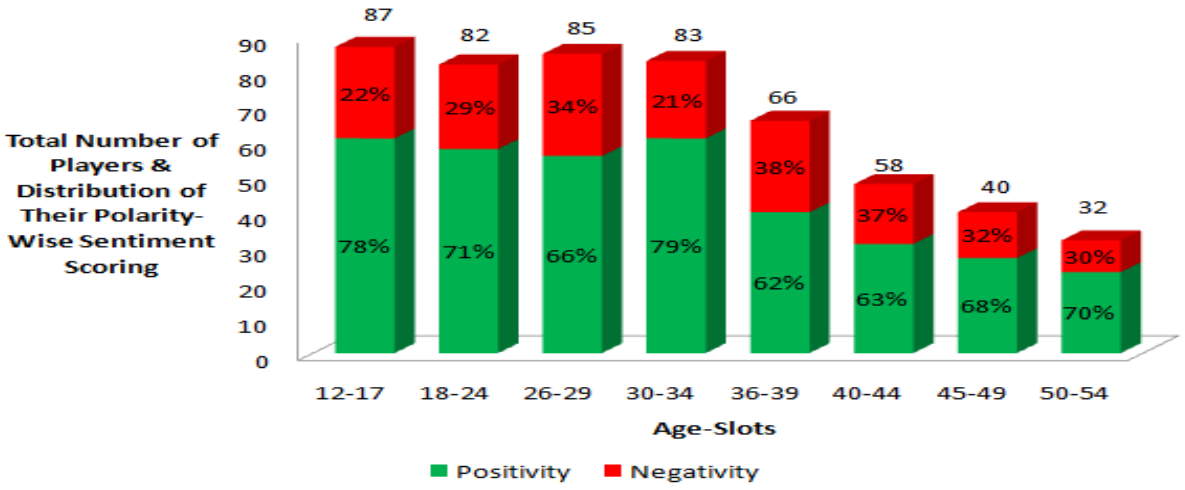


Figure 1.6: Senti-Mentality Age Wise

1.4.1.3 Gender Wise

Senti-mentality also changes with gender as reported in the following Figure 1.7. Although the distribution of male and female players is not even and the lexicon is always updated, but it has been observed that woman are more positive than man!

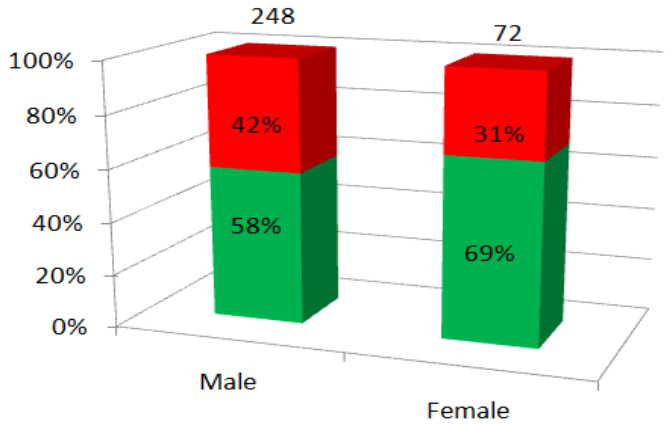


Figure 1.7: Senti-Mentality Gender-Wise

1.4.1.4 Other-Wise

Some of the important dimensions that have not been explored so far include country, city, profession etc. Combinations of the various dimensions such as location-age, location-profession, sex-wise, language-location may reveal some interesting study.

1.5 Evaluation of the Generated Resources

Andera Esuli and Fabrizio Sebastiani (Esuli and Fabrizio, 2006), the inventors of the SentiWordNet, have calculated only the reliability of the sentiment scores attached to each synset in the SentiWordNet. They have tagged sentiment words in the English WordNet with positive and negative sentiment scores. Two extrinsic evaluation strategies have been proposed for the developed SentiWordNet(s) based on the two quality measures: **coverage** and **credibility** (Das and Bandyopadhyay, 2010(c));(Das and Bandyopadhyay, 2010(e)). The evaluation has been done only on the SentiWordNet developed for Bengali.

1.5.1 Coverage

To evaluate the coverage experiments have been carried out with NEWS and BLOG corpora for subjectivity detection (discussed in Chapter 2). Sentiment lexicons are generally domain independent and it provides a good baseline while working with sentiment analysis systems. The coverage of the developed Bengali SentiWordNet is evaluated by using it in a subjectivity classifier (Das and Bandyopadhyay, 2009(a)). The statistics of the NEWS and BLOG corpora is reported in Table 1.10.

	NEWS	BLOG
Total number of documents	100	-
Total number of sentences	2234	300
Avgerage number of sentences in a	22	-
Total number of wordforms	28807	4675
Avgerage number of wordforms in a	288	-
Total number of distinct wordforms	17176	1235

Table 1.10: Statistics of Bengali Corpus, used to measure the Coverage of the developed SentiWordNet (Bengali)

Languages	Domain	Precision	Recall
English	**MPQA**	76.08%	83.33%
	IMDB	79.90%	86.55%
Bengali	**NEWS**	72.16%	76.00%
	BLOG	74.6%	80.4%

Table 1.11: Comparative Evaluation of a Subjectivity Classifier using SentiWordNet (English) and SentiWordNet (Bengali)

For comparison with the coverage of English SentiWordNet the same subjectivity classifier (Das and Bandyopadhyay, 2009(a)) has been applied on Multi Perspective Question Answering (MPQA) (NEWS) and IMDB Movie review corpus using the English SentiWordNet. The result of the subjectivity classifier on both the corpus proves that the coverage of the Bengali SentiWordNet is reasonably good and comparable. The experiments on the IMDB corpus have yielded high precision and recall scores and the subjectivity word list used in the subjectivity classifier is developed from the IMDB corpus. Hence the

developed Bengali SentiWordNet is domain independent and its coverage is very good as shown in Table 1.11. The SentiWordNet(s) for other languages have not been evaluated as there is no publicly available sentiment annotated data for these languages.

1.5.2 Credibility

This evaluation metric measures the reliability of the associated polarity scores in the sentiment lexicons. A typical approach to sentiment analysis is to start with a lexicon of positive and negative words and phrases. In the lexicon, entries are tagged with their prior out of context polarity. To measure the reliability of polarity scores in the developed Bengali SentiWordNet, a polarity classifier (Das and Bandyopadhyay, 2010(h)) (discussed in Chapter 3) has been developed based on the Bengali SentiWordNet along with some other linguistic features.

Features	Overall Performance
SentiWordNet	47.60%
SentiWordNet + Negative Word	50.40%
SentiWordNet + Negative Word + Stemming Cluster	56.02%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word	58.23%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word Parts Of Speech	61.9%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech +Chunk	66.8%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech + Chunk +Dependency tree feature	70.04%

Table 1.12: Performance of a Polarity Classifier Using Bengali SentiWordNet (Bengali) by Feature Ablation

Feature ablation method shows that the associated polarity scores in the developed Bengali SentiWordNet are reliable. Table 1.12 shows the performance of a polarity classifier using the Bengali SentiWordNet. The polarity wise overall performance of the polarity classifier is reported in Table 1.13.

Polarity	Precision	Recall
Positive	56.59%	52.89%
Negative	75.57%	65.87%

Table 1.13: Polarity-wise Performance of a Polarity Classifier Using SentiWordNet (Bengali)

Comparative study with an English polarity classifier that works only with prior polarity lexicon is necessary but no such works have been identified from the literature.

300 words have been arbitrarily chosen from the developed Hindi SentiWordNet for human evaluation. Two persons have been asked to manually check these words and the result is reported in Table 1.14. The coverage of the Hindi SentiWordNet has not been evaluated, as no manually annotated sentiment corpus is available.

Polarity	Positive	Negative
Percentage	88.0%	91.0%

Table 1.14: Evaluation of Assigned Polarity Scores for Developed SentiWordNet (Hindi)

For Telugu we rely on the Dr. Sentiment with Telugu words on screen. Only 30 users have played the Telugu language specific game till date. Total 920 arbitrary words have been tagged and the accuracy of the polarity scores is reported in Table 1.15. The coverage of the Telugu SentiWordNet has not been evaluated, as no manually annotated Telugu sentiment corpus is available.

Polarity	Positive	Negative
Percentage	82.0%	78.0%

Table 1.15: Evaluation of Assigned Polarity Score of Developed SentiWordNet (Telugu)

1.6 Expected Impact of the Resource

Undoubtedly the generated SentiWordNet(s) are important resources for sentiment/opinion or emotion analysis task. Moreover the other non linguistic psychological dimensions are very much important for further analysis in several newly discovered sub-disciplines such as: Geospatial Information retrieval (Egenhofer, 2002), Personalized search (Gaucha et al., 2003) and Recommender System (Adomavicius and Tuzhilin, 2005) etc.

Deciding on the data structure for storing the SentiWordNet is not trivial. Presently RDBMS (Relational Database Management System) has been used. Several tables are used to keep user's clicking log and their personal information.

There is a pertinent question on the reliability of the word level Google translation API. It is well accepted that word sense disambiguation (WSD) is a big problem and a separate research issue in NLP and sentiment lexicon is not an exception. Surely there will be errors in word level Google translation API and thus further cleaning is required. Assignment of Sense ID to each synset is the future research aspect which has been further discussed in the Conclusion chapter.

Publications

1. **Amitava Das** and Sivaji Bandyopadhyay. 2011. ***Dr Sentiment Knows Everything!*** *In the Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011 Demo Session)*, Pages 50-55, Portland, Oregon, USA.
   http://www.aclweb.org/anthology/P11-4009

2. **Amitava Das.** 2011. ***PsychoSentiWordNet***. *In the Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011 Student Session)*, Pages 52-57 Portland, Oregon, USA.
   http://aclweb.org/anthology/P/P11/P11-3010.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2010(c). ***Dr Sentiment Creates SentiWordNet(s) for Indian Languages Involving Internet Population***. *In the Proceeding of IndoWordNet Workshop, ICON 2010*, Kaharagpur, India.
   http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/14_iwn_SentiWordNet%28s%29%20for%20Indian%20Languages.pdf

4. **Amitava Das** and Sivaji Bandyopadhyay. 2010(d). ***Towards The Global SentiWordNet***. *In the Proceeding of the Workshop on Model and Measurement of Meaning (M3), PACLIC 24th*, Pages 799-808, Sendai, Japan.
   http://www.amitavadas.com/Pub/GSN.pdf

5. **Amitava Das** and Sivaji Bandyopadhyay. 2010(e). ***SentiWordNet for Indian Languages***. *In the Proceeding of the 8th Workshop on Asian Language Resources (ALR 8), COLING 2010*, Pages 56-63. Beijing, China.
   http://aclweb.org/anthology/W/W10/W10-3208.pdf
   http://59.108.48.12/proceedings/coling/coling2010/ALR/pdf/ALR08.pdf

6. **Amitava Das** and Sivaji Bandyopadhyay. 2010(f). ***SentiWordNet for Bangla***. *In the Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary (KSE4)*, Mysore, India.
   http://www.ldcil.org/up/conferences/Dictionary/Dict-Event_Schedule.pdf
   http://www.amitavadas.com/Pub/SentiwordNet%20%28Bengali%29.pdf

Chapter 2

Sentiment / Subjectivity Detection

Sentiment analysis task seeks to analyze expressed sentiment within a piece of text, be it a word, a phrase, a sentence or a document. The overall sentiment analysis task is mainly divided into two main subtasks: ***sentiment detection*** and ***sentiment classification***. The role of sentiment detection task is to detect the presence of sentiment/opinion in any piece of text while the sentiment classification task concentrates on the classification of those texts according to their semantic orientation (positive or negative or any further fine-grained emotion classes). Sentiment detection is well known with the terminology ***Subjectivity Detection*** within the research community. The various subjectivity detection methods proposed in the present work have been reported in this chapter.

The rest of the chapter is organized as follows. The definition of subjectivity is discussed in the section 2.1. Sentimental behavior is the basic reflection of human intelligence. Sentiment detection in written text is challenging even for human beings. Therefore, computational subjectivity detection is a challenging and enigmatic research problem. The challenges are described with some specific examples in the section 2.2. A number of research endeavors could be found in the literature that has attempted subjectivity detection from natural language text for various domains and languages. These are discussed in the section 2.3. The experiments on subjectivity detection in the present work have been carried out for both English and Bengali languages. The corpus development process has been described in the section 2.4. Feature Engineering involves feature identification and extraction and is the most crucial issue for any kind of NLP applications. The subjectivity detection task started with a large set of features and finally concluded with a selected list of linguistically extractable lexical, syntactic and discourse-level feature set, which best contributes for both English and Bengali languages. The details of the feature extraction technique are described in the section 2.5. A series of experiments have been conducted starting from the rule based **Theme Detection** technique and followed by machine learning techniques like **Conditional Random Field** (CRF) and **Genetic Algorithm**. The details of these techniques are elaborated in the section 2.6.

2.1 What is Subjectivity?

Any scientific research needs to know the proper definitions of the problems in order to solve it. The essential questions that should be answered at the beginning of the sentiment analysis research includes "***What is sentiment or opinion?***" and "***What are the emotional data we need to extract from the textual data for the sake of any real time usability?***". Various research endeavors try to answer these questions in the light of psychology, philosophy and psycholinguistics and even with cognitive science. The researchers attempted to give their own definitions. Among those research endeavors the General Inquirer (1966) (Stone, 1966) System and the Subjectivity definition by Janyce Wiebe (Wiebe et. al., 1990) are the milestones that mark the avenue to the current research trend of today.

Chapter 2 Sentiment / Subjectivity Detection

The sentiment analysis research started as a content analysis research problem in the behavioral science. The General Inquirer System[1] (1966) (Stone, 1966) is the first attempt in this direction. The aim was to gain understanding of the psychological forces and perceived demands of the situation that were in effect when the document was written. The system usually counts the occurrences of positive or negative emotion instances in any particular piece of text. Although the sentiment analysis research has started long back but the question "***What is sentiment or opinion?***" remain unanswered till date! Moreover, no complete set of psychological forces could be defined that really affect the sentiment of the writer.

"***How the mind works is still a mystery. We understand the hardware, but we don't have a clue about the operating system.***"

James Watson (Nobel laureate)

With the advances of World Wide Web (WWW) and other digital technologies, text have become the primary medium of representing and transmitting information, as evidenced by the pervasiveness of e-mails, instant messages, documents, weblogs, news articles, homepages and printed materials. Modern lives have become saturated with electronic text information, and there is an increasing urgency to develop technology to manage and make sense of the resulting information overload. While keyword-based and statistical approaches have enjoyed some success in assisting information retrieval, data mining, search engines and natural language processing (NLP) systems, there is a growing necessity to develop computational sentimental intelligence. The Sentiment Analysis research makes its own way to develop systems that can organize, detect, classify and analyze human opinion from electronic text.

"***What other people think***" has always been an important piece of information for most of us during any decision-making process. For example, we try to know other people's opinion about any electoral candidate before casting our vote or try to collect feedback of others regarding any product before buying it. General search engines or Information Retrieval systems does not have these facilities. To meet the end user's needs Janyce Wiebe came up with her innovative idea of ***subjectivity***. The term ***Subjectivity*** can be defined as the ***Topical Relevant Opinionated Sentiment***. Let us take a look at the following example:

`I spend my holiday with my new Sony CyberShot in Queensland, a splendid place. The camera makes my holiday special!`

If the above text is treated as a review then anyone, who is interested to buy the same camera will look into it to discover how a former user recommends it. The person may not be interested about the writer's sentiment about the place, rather will be interested in topical relevant opinion about the camera! The topical relevancy can be found only in the second sentence. The first sentence also depicts sentiment about the place where the writer spends his/her holiday with the phrase "***splendid***

[1] http://www.wjh.harvard.edu/~inquirer/

place" but it is not relevant regarding the camera. The ***Topical Relevant Opinionated Sentiment*** detection is well known as ***Subjectivity Detection.*** Janyce Wiebe borrowed the definition of opinion from a Psycholinguistics research which states: ***an opinion could be defined as a private state that is not open to objective observation or verification*** (Quirk et. al., 1985).

An opinion could be defined as a private state that is not open to objective observation or verification.

(Quirk et al., 1985)

2.2 Subjectivity Detection: the Challenges

Subjectivity detection is a challenging research problem. Sentiment is one of the finest aspects of natural human intelligence but still ***Topical Relevant Opinionated Sentiment*** detection or ***Subjectivity*** detection is not trivial even for human intelligence. Let us elaborate with some relevant examples.

Type: Product Review My camera broke in two days.

Detection of subjectivity from the previous sentence is very ambiguous. The example sentence reports an incident but the reasons are not specified. Was the quality of the camera too bad or any accident took place? The issue could be resolved by additional discourse level information from the preceding as well as succeeding sentences. If the following sentence is the successor then probably the previous sentence is not a subjective expression.

I felt down from stairs and thus I lost it.

On the contrary if the following sentence is the successor in the discourse the previous sentence is depicting a negative opinion.

I am very upset and decided to buy my next camera, should have long life rather than plenty of features.

Therefore, subjectivity is not only syntactic or semantic issue of a language but pragmatics is also involved. Let us look at another example sentence on movie review.

Type: Film Review, Film Name: Deep Blue Sea, Holder: Arbitrary-outside of theatre

Oh, this is blue!

The sentence is an example of positive subjective expression as well as a metaphor. Subjectivity detection in metaphors is a very complex issue. In the sentence, there is no opinionated word or subjective marker and the only clue is the word "***blue***". The metaphor is quite contextual as the word is taken from the movie name itself.

Detection of sentiment from only written text is very challenging as there is no clue of emphasizing or other tonal changes that are encoded in the spoken language. Let us look at the following example:

I will not go for shopping with you today.

Apparently there is no sentiment in the previous sentence that rather appears to be a factual sentence. In spoken language the scenario could be different. In the following example sentences, the bold faced words represent the emphasize during the speech act. The examples show how emphasize in the speech act changes the subjectivity and the aspect of subjectivity.

I will not go for shoppingwith you today.

I will not go for shopping with **you** today.

I will not go for shopping with you **today.**

I will not go for **shopping** with you today.

Therefore sentiment of human beings depends on many factors. We are empowered with our natural intelligence to act on these sentiments but computers are emotionally challenged.

2.3 Previous Studies

In the year of 1999, Jaynce Wiebe (Wiebe et. al, 1999) defined the term Subjectivity in Information Retrieval perspective. Sentences are categorized into two genres as Subjective and Objective. Objective Sentences are used to objectively present factual information and subjective sentences are used to present opinions and evaluations.

Researchers have experimented with several methods to solve the problem of subjectivity detection using SentiWordNet, Subjectivity Word List etc. as prior knowledge database. But subjectivity detection is a domain dependent and context dependent problem (Aue and Gamon, 2005). Hence building a prior knowledgebase for subjectivity detection will never be exhaustive. Moreover, Sentiment/opinion changes its polarity orientation with time. Hence, subjectivity detection needs a most sophisticated algorithm to capture and effectively use the sentiment pragmatic knowledge. The algorithm should be customizable for any new domain and language.

Previous works in subjectivity identification have helped in the development of a large collection of subjectivity clues. These clues include words and phrases collected from manually developed annotated resources. The clues from manually developed resources include entries from adjectives manually annotated for polarity (Hatzivassiloglou and McKeown, 1997), and subjectivity clues listed in (Wiebe, 1990). Clues learned from annotated data include distributionally similar adjectives and verbs (Wiebe, 2000) and n-grams (Dave et. al, 2003). Low frequency words are also used as clues. Such words are informative for subjectivity identification.

An opinion could be defined as a private state that is not open to objective observation or verification (Quirk et. al., 1985). Opinion extraction, opinion summarization and opinion tracking are three important techniques for understanding opinions. Opinion-mining of product reviews, travel advice, consumer complaints, stock market predictions, real estate market predictions, e-mail etc. are areas of interest for researchers since last few decades.

Most research on opinion analysis has focused on sentiment analysis (Wiebe, 1990) and subjectivity detection (Wiebe, 2000; Dave et. al, 2003). Methods on the extraction of opinionated sentences in a structured form can be found in (Aue and Gamon, 2005). Some machine learning text labeling algorithms like Conditional Random Field (CRF) (Zhao et. al, 2008), Support Vector Machine (SVM) (Hatzivassiloglou and Wiebe, 2000) have been used to cluster same type of opinions. Application of machine-learning techniques to any NLP task requires a large amount of data. It is time-consuming and expensive to hand-label the large amount of training data necessary for good performance. Hence, use of machine learning techniques to extract opinions in any new language may not be an acceptable solution.

Opinion analysis of news documents is an interesting area to explore. Newspapers generally attempt to present the news objectively, but textual affect analysis in news documents shows that many words carry positive or negative emotional charge (Chesley et. al, 2006). Some important works on opinion analysis in the newspaper domain are found in (Nasukawa and Yi, 2003) but no such efforts have been taken up in Indian languages especially in Bengali. Development of Subjectivity Classifier for a new language demands sentiment lexicon and gold standard annotated data for machine learning and evaluation.

(Mihalcea et. al., 2007) have proposed several techniques including translation methodology to develop Subjectivity resources in cross-lingual perspective for Romanian language from English. The main problem faced during the translation process is the presence of inflected words that require stemming as a solution.

In case of sentence level subjectivity annotation a parallel corpus based approach has been proposed in (Mihalcea et. al., 2007). But for Indian languages especially Bengali it is very hard to collect appropriate parallel corpora. (Wan, 2008) has proposed generation of Chinese reviews from English texts by Machine Translation. Publicly available tools like GoogleTrans[2], Yahoo Babel Fish and a word level translation module have been used. When we started, there was no publicly available Machine Translation system for English-Bengali language pair (as of this time, an English-Bengali GoogleTrans is available) though an English-Hindi machine translation system was available in GoogleTrans. In the present work, rule based sentence level subjectivity annotation has been done that is d finally manually checked for validation.

[2] http://translate.google.com/

Another significant effort on subjectivity annotation is found in (Anthony et. al., 2005). Opinion/Sentiment mining is identified as a very domain specific problem. The problem of unavailability of large amount of labeled data for fully supervised learning approaches has been addressed. Hence the proposed solution in (Anthony et. al., 2005) is a Subjectivity classifier, customizable to any new domain. The aim of the present work is to devise a general architecture for developing a Subjectivity classifier for a new language with domain and language dependency. There are other research activities with multiple domains, e.g., (Godbole et. al., 2007).

As a growing number of people use the Web as a medium for expressing their opinions, the Web is becoming a rich source of various opinions in the form of product reviews, travel advice, social issue discussions, consumer complaints, movie review, stock market predictions, real estate market predictions, etc. Present computational systems need to extend the power of understanding the sentiment/opinion expressed in an electronic text. The topic-document model of information retrieval has been studied for a long time and several systems are available publicly since last decade. On the contrary, Opinion Mining/Sentiment Analysis is still an unsolved research problem. Although a few systems like Bing, Twitter Sentiment Analysis Tool etc. are available in the web since last few years, still more research efforts are needed to match the user satisfaction level and social need.

2.4 Corpus in the Present Work

All the subjectivity detection experiments in the present work are adopted and tested for both English and Bengali. In case of English, experiments have been carried out with the domain corpus from the *news* and the *review* domains. The most popular Multi Perspective Question Answering (MPQA)[3] corpus (Wiebe and Riloff, 2006) has been chosen for news domain and the International Movie Database (IMDB)[4] (Pang et al., 2002) has been chosen for the review domain.

In the MPQA corpus the **private states** in a sentence are annotated at phrase/expression level but there is no sentence level subjectivity annotation. Thus a semi-automatic technique has been proposed for sentence level subjectivity annotation and the process is described in the Section 2.4.1. The details of the IMDB movie review corpus are discussed in the section 2.4.2. There were no opinion / sentiment corpus available for Bengali. Sentiment annotated Bengali corpus has been developed for the news and blog domains. The Bengali news corpus have been collected from the editorial pages, i.e., Reader's opinion section or Letters to the Editor Section, from the web archive of a popular Bengali newspaper and then manually annotated. The overall process of the Bengali news corpus development is described in Section 2.4.3. A random collection of 123 blog posts containing a total of 12,149 sentences have been retrieved from the Bengali web blog archive[3] (especially from comics, politics, sports and short stories) to prepare the Bengali blog corpus.

[3] http://www.cs.pitt.edu/mpqa/
[4] http://www.cs.cornell.edu/People/pabo/movie-review-data/

2.4.1 Semi-Automatic Subjectivity Annotation for MPQA

The annotation scheme of the MPQA corpus is designed to identify the key components and properties of various types of opinions / emotions / sentiments / speculations / evaluations and other private states. The properties of a private state frame include the source of the private state (i.e., whose private state is being expressed), the target (i.e., what the private state is about) and various other details involving intensity, significance and type of attitude. The annotation within the MPQA corpus is not at sentence level but at word or expression level. Every sentence in a corpus does not express sentiment / opinion. The sentences that express topical relevant sentiments / opinions are identified as **subjective** sentences and the sentences that report about any fact or incident are defined as **factual** sentences. Private states of subjective expressions are classified into two basic categories, i.e., **direct subjective** frames and **express subjective** frames. **Objective speech event** frames have been defined to distinguish opinion-oriented materials from other factual materials.

Direct Subjective Frame: A private state containing a direct subjective element is called a direct subjective frame. For example, in the sentence the word '**fears'** represents a private state and is annotated as **Direct Subjective Frame**.

"The U.S. **fears** a spill-over," said Xirao-Nima.

Expressive Subjective Frame: A private state containing no direct opinion but only subjective references to opinion is called an expressive subjective frame. For example, in the sentence the phrase "**full of absurdities**" represents a private state and is annotated as **Expressive Subjective Frame**.

"The report is full of absurdities," Xirao-Nima said.

Objective Speech Event: This is purely the factual part of any event. For example, the following sentence does not carry any opinionated information but is a description of a fact or event.

O'Leary said "the incident took place at 2:00pm."

A semi-automatic method has been adopted to annotate the subjective sentences of MPQA. The hypothesis is that if a sentence has any Direct Subjective or Expressive Subjective expression then the sentence must be subjective itself. Thus, the sentences containing either of the two private states (i.e., Direct Subjective and Expressive Subjective) are extracted as subjective and the sentences containing only objective speech event or no annotated private states are discarded. But there is a problem with this semi-automatic method. As stated earlier, subjectivity refers to the Topical Relevance Sentiment. Thus some sentences may have sentiment but they may have no contextuality with the main topic of the document. These non-relevant sentimental sentences are then manually discarded. Almost 12% sentences have been manually discarded. To do this manual checking, a simple tool, called MPQA Explorer, has been developed that highlights the private stases, i.e., direct subjective (RED), express subjective (CYAN) and objective speech events (YELLOW) with different colors. A snapshot of the MPQA Explorer tool is shown in Figure 2.1.

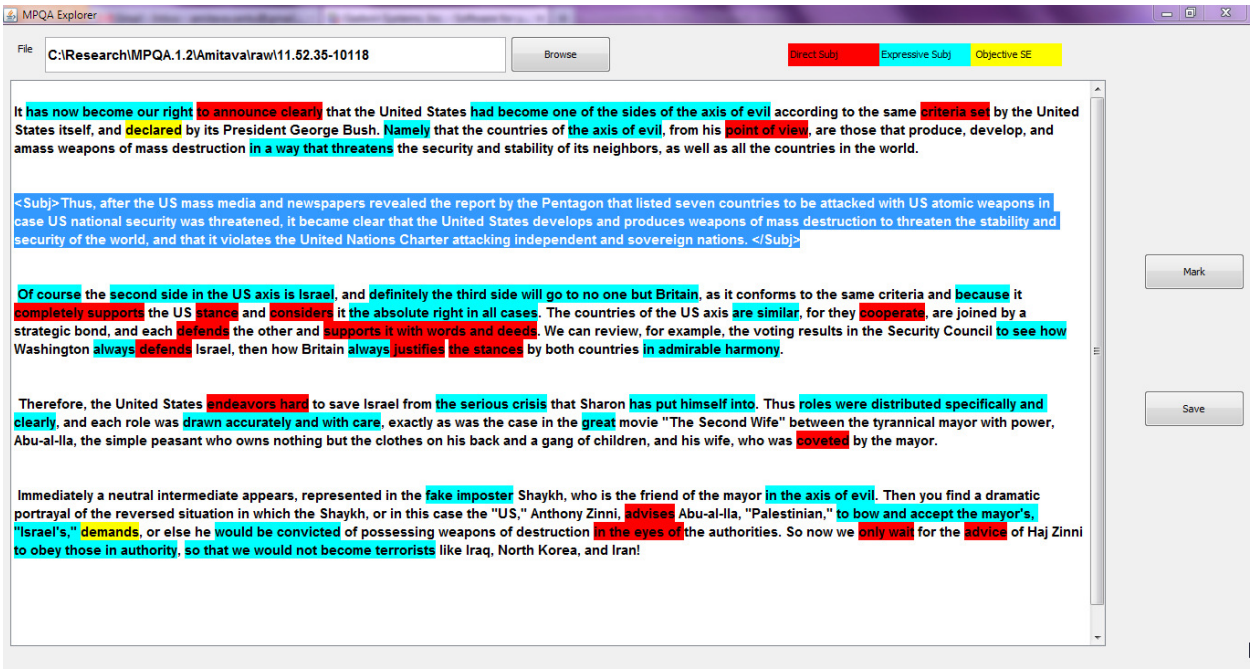


Figure 2.1: The MPQA Explorer

The experiments started with MPQA version 1 (released in 2003) where there are no subjective sentence marking. MPQA version 2 (released in late 2008) is annotated for the subjective sentences. The accuracy of the proposed semi-automatic subjectivity marking has been calculated with respect to the MPQA version 2 and approximately 90% accuracy (Precision: 100 and Recall: 81.8. F-Score: 89.98) has been obtained. If the accuracy is automatically calculated, i.e., without the manual discarding of the 12% sentences, then the accuracy figure is as low as 78.3% (Precision: 98.0 and Recall: 65.2. F-Score: 78.3). The improved accuracy figures prove the necessity of human involvement. Both version of MPQA are downloadable from the website: http://www.cs.pitt.edu/mpqa/mpqa_corpus.html.

2.4.2 English IMDB Movie Review Corpus

There are various types of review corpus available in the web. The reviews can belong to any genre like movie review, product review, tourism review, electoral review etc. The review corpus is saturated with rich textual sentimental information and has generated high interests for the sentiment analysis researchers. For the experiments in the present work, the movie reviews corpus[5] developed by (Pang and Lee, 2005) has been identified. The data source was the Internet Movie Database (IMDb). Only the reviews where the author rating was expressed either with stars or with some numerical value (other conventions varied too widely to allow for automatic processing) have been selected. Ratings were automatically extracted and converted into one of three categories: positive, negative or neutral. The corpus consist of 752 negative and 1301 positive reviews, with a total of 144 non-prolific reviewers and

[5] http://www.cs.cornell.edu/people/pabo/movie-review-data/

an average 20 reviews per reviewer. The corpus is maintained at sentence level with sentence level subjectivity marked. Thus, no subjectivity annotations are required unlike the MPQA.

2.4.3 NEWS and BLOG Sentiment Corpora in Bengali

To check the cross domain efficiency and compatibility of the proposed subjectivity detection algorithms and lexicon resources, corpora of different domains are required. The NEWS (i.e., MPQA) and the IMDB Movie review database have been identified for English while the NEWS and the BLOG corpora for Bengali have been developed as there is no corpus available for sentiment analysis in Bengali.

The Bengali NEWS corpus has been collected from a popular news web archive and then further manually annotated. A focused web crawler (Ekbal and Bandyopadhyay, 2008) retrieves the web pages in Hyper Text Markup Language (HTML) format from the news archive of a leading Bengali news paper within a range of dates provided as input. The news documents in the archive are stored in a particular fashion. The user has to give the range of dates as starting and ending as **yy-mm-dd** format. The crawler generates the Universal Resource Locator (URL) address for the index (first) page of any particular date. The index page contains actual news page links and links to some other pages (e.g., Advertisement, Editorial, TV schedule, Tender, Comics and Weather etc.) that do not contribute to the corpus generation. The HTML files that contain news documents are identified and the rest of the HTML files are not considered further. The Bengali texts in the archive are written in dynamic fonts and the Bengali pages are generated on the fly on the screen, i.e., only when the system is online and is connected to the web. Moreover, the newspaper archive uses graphemic coding whereas orthographic coding is required for text processing tasks. Hence, Bengali texts, written in dynamic fonts are not suitable for text processing activities. In graphemic coding, a word is coded according to the constituent graphemes. But in orthographic coding the word is coded according to the constituent characters. Conjuncts have separate codes in graphemic coding while these are coded in terms of the constituent consonants in orthographic coding. A code conversion routine has been written to convert the dynamic codes used in the HTML files to represent Bengali text to Indian Standard Code for Information Interchange (ISCII) codes. A separate code conversion routine has been developed for converting ISCII codes to UTF-8 codes. For the rest of the experiments the standard and universal UTF-8 encoding has been used.

From the collected document set, some documents from Letters to the Editor Section have been chosen for the annotation task. Documents that appeared within an interval of four months are chosen on the hypothesis that these letters to the editors will be on related topics. A simple annotation tool has been designed for annotating the subjective sentences. A snapshot of the tool has been shown in the Figure 2.2. The tool highlights the sentiment words (based on the occurrence of the word in the SentiWordNet (Bengali), (described in Chapter One) by two different colors within a document according to their sentiment orientation categories (GREEN: Positive words, RED: Negative words as reported in the Figure 2.2). The tool also highlights the title words (YELLOW) and theme words (BLUE), automatically identified by the rule-based theme detection technique (described in section 2.6.1). For example, the words "নরেন্দ্র মোদি" and "ঘুষ" are the title words as they occur in the title of the document and have been highlighted

in yellow. Words like "গুজরাট" and "মামলার" are the theme words highlighted in blue. The words highlighted in either green ("অভিযোগ" and "প্রভাবিত") or red ("স্বীকার") are the sentiment words extracted from SentiWordNet (Bengali).

Figure 2.2: The Subjectivity Annotation Tool for Bengali

Three annotators participated in the present task. The documents with such annotated sentences are saved in XML format. The XML tag "<SS>" stands for subjective sentence as shown in the Figure 2.3.

The complete manual development of the annotated corpus would be expensive. Therefore, a semi-automatic annotation technique has been followed. In the present work, a rule based sentence level subjectivity annotation (described in section 2.6.1) has been done for Bengali that is finally checked for validation by the same three human annotators. The present technique is relatively simple and less human interactive that can be followed for any new language with limited number of resources. This technique helps to increase the speed of the annotation process. Finally, 100 annotated documents have been produced. Some statistics about the Bengali news corpus is represented in the Table 2.1.

<document docid="Modi-324" encoding="UTF-8">

<header>

<title>২০০২ সালের দাঙ্গার ঘটনায় আইনজীবীদের ঘুষ দিয়েছিলেন নরেন্দ্র মোদি </title>

</header>

<body>

<Sentence id="1">

<SS>ভারতের গুজরাট রাজ্যের মুখ্যমন্ত্রী নরেন্দ্র মোদি ২০০২ সালের দাঙ্গার ঘটনায় একটি মামলার বাদীর আইনজীবীদের ঘুষ দিয়েছিলেন বলে অভিযোগ উঠেছে। </SS>

</Sentence>

<Sentence id="6">

<SS>ঘুষ দিয়ে ওই মামলার শুনানি প্রভাবিত করাই ছিল তাঁর উদ্দেশ্য।</SS>

</Sentence>

<Sentence id="7">

দাঙ্গার ঘটনা তদন্তে গঠিত নানাবতী কমিশনে দেয়া গোয়েন্দা কর্মকর্তা শ্রীকুমারের নথির একটি কপিও সাংবাদিকদের কাছে সরবরাহ করেন মল্লিকা।

</Sentence>

<Sentence id="8">

<SS>তিনি দাবি করেন, আরেক গোয়েন্দা কর্মকর্তা সঞ্জীব ভাট গত মে মাসে নানাবতী কমিশনে হাজির হয়ে ঘুষের বিষয়টি স্বীকার করেন। </SS>

</Sentence>

Figure 2.3: Bengali Corpus Subjectivity Annotation Scheme

	NEWS	BLOG
Total number of documents	100	-
Total number of sentences	2234	300
Avgerage number of sentences in a document	22	-
Total number of wordforms	28807	4675
Avgerage number of wordforms in a document	288	-
Total number of distinct wordforms	17176	1235

Table 2.1: Statistics of Bengali Corpus developed for Subjectivity Detection

A small Bengali BLOG corpus has been collected and manually annotated. Random collection of 123 blog posts containing a total of 12,149 sentences are retrieved from the Bengali web blog archive[6] (especially from comics, politics, sports and short stories) to prepare the corpus. It has been noticed during annotation that subjectivity annotation for BLOG corpus is more trivial than NEWS corpus as people generally express their opinion/ sentiment directly in the BLOG compared to the NEWS text. A brief statistics about the corpus has been reported in the Table 2.1.

2.5 Learning Subjectivity Clues through Feature Engineering

Feature engineering involves feature identification and feature extraction. It plays a crucial role in any kind of NLP task. In the subjectivity detection task, the aim is to find out the most effective and concise set of features that work across various languages and domains. Features are the linguistic clues to detect the desired pattern and the clues may exist at any level like lexical, syntactic or discourse level. As subjectivity refers to the topical relevant sentiment, therefore a system needs to know the theme of any piece of text to detect the presence of subjectivity in that text. The complete list of lexical, syntactic and discourse level feature sets are reported in the Table 2.2.

Types	Features
Lexical	POS
	SentiWordNet
	Frequency
	Stemming
Syntactic	Chunk Label
	Dependency Parsing
Discourse Level	Title of the Document
	First Paragraph
	Average Distribution
	Theme Word

Table 2.2: Features for Subjective Detection

Once the best feature set has been identified then the challenge is to extract those features effectively. Various linguistics tools that are used to extract the features for both the languages are reported in the following sub-sections.

2.5.1 Lexical Features

Lexical analysis plays a crucial role to identify sentiments from a text. For example, words like ***love***, ***hate***, ***good*** and ***favorite*** directly indicate sentiment. Therefore, we need to extract the basic lexical clues in order to identify the subjectivity. The lexical features that are used in subjectivity detection are part of

[6] www.amarblog.com

speech (POS) category, sentiment words from SentiWordNet, frequency and the stemmed root of a word.

2.5.1.1 Part of Speech (POS)

It has been identified in a number of research activities (Hatzivassiloglou and Wiebe, 2000; Chesley et. al., 2006) that sentiment bearing words are mainly adjective, adverb, noun and verbs.

The Stanford Parser[7] has been used for identifying the POS tags in case of English text. The Bengali shallow parser[8] developed under the project Indian Languages to Indian Languages Machine Translation Systems (IL-ILMT) has been used. The project is funded by Department of Information Technology, Government of India and is being executed by a consortium of 14 institutes.

2.5.1.2 Sentiment Words

The dictionary based approach is very standard for sentiment analysis. Several prior polarity sentiment lexicons are available for English: SentiWordNet (Esuli et. al., 2006), Subjectivity Word List (Wilson et. al., 2005), WordNet Affect list (Strapparava et al., 2004) and Taboada's adjective list (Taboada et al., 2006). SentiWordNet and Subjectivity Word List have been identified as the most reliable lexicons. The SentiWordNet is widely used and the Subjectivity Word List is robust in terms of performance. Taboada's adjective list is not considered in the present work as it contains only 1,719 adjectives. WordNet Affect list is also not considered as it contains emotion information. Words that are present in the SentiWordNet carry sentiment information. A sentiment lexicon has been generated from both the SentiWordNet and the Subjectivity Word List after removing the duplicates and applying other filtering techniques as described in the Chapter one. The generated merged lexicon is used for the subjectivity detection task. The SentiWordNet (Bengali)[9] as described in the Chapter one is used for the present task.

These features are individual sentiment words or word n-grams (multiword entities) with strength measure as strong subjective or weak subjective. Strong and weak subjective measures are treated as a binary feature in the rule based and supervised classifiers. Words which are collected directly from the SentiWordNet are tagged with positivity or negativity scores. The subjectivity score of these words are calculated as:

$$E_s = | S_p | + | S_n |$$

where E_s is the resultant subjective measure and S_p, S_n are the positivity and negativity scores respectively. The words with subjectivity score greater than 0.4 are considered and all other words are discarded (Wilson et. al., 2005).

[7] http://nlp.Stanford.edu/software/lex-parser.shtml
[8] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php
[9] http://www.amitavadas.com/sentiwordnet.php

2.5.1.3 Frequency

Frequency of a word within a discourse determines its importance as informative in order to understand the theme of the discourse. Frequency does not play any role to understand the sentiment but as subjectivity refers to topical relevance sentiment, therefore frequency helps to identify topical relevance. If one or more high frequent word co-occurs with a sentiment word then the sentence may have subjectivity. After removal of function words and POS annotation, the system generates four separate high frequent word lists for the four POS categories: Adjective, Adverb, Verb and Noun. Word frequency values are effectively used as a crucial feature in the Subjectivity classifier.

2.5.1.4 Stemming

Several words in a sentence that carry sentiment information may be present in inflected forms. Stemming is necessary for such inflected words before they can be searched in the appropriate lists. Due to non availability of a good Bengali stemmer, a stemming cluster technique based Bengali stemmer (Das and Bandyopadhyay, 2010(l)) has been developed. The stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center. The Porter Stemmer[10] has been used for English.

2.5.2 Syntactic Features

Syntactic relations affect the overall sentiment expressed by a piece of text. Let us consider the following example sentence:

```
(The steering) (is) (not predictable).
```

In the previous example, the evaluative expression "***not predictable***" depicts the sentiment. But does this sentence depict subjectivity? To answer this we need to go further at syntactic level. The "***not predictable***" phrase ultimately relates the subject of the sentence, i.e., "the steering", which may or may not have topical relevance in any discourse. Therefore, the following thumb rule is applied: if any sentiment word/phrase has any syntactic/dependency relation with any theme words then only the sentence is considered as subjective. The following features are identified as relevant.

2.5.2.1 Chunk Label

Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. The Stanford Parser has been used for identifying the chunk labels in English. For Bengali the shallow parser developed under the project Indian Languages to Indian Languages Machine Translation Systems (IL-ILMT) has been used. Chunk boundaries help to understand the local dependencies between words such as, theme words and sentiment words.

[10] http://tartarus.org/~martin/PorterStemmer/

2.5.2.2 Dependency Parser

Dependency feature is very useful to identify intra-chunk polarity relationship. It is very often a language phenomenon that modifiers or negation words are generally placed at a distance with evaluative polarity phrases. The Stanford Dependency Parser for English has been used in the present work. But unfortunately no dependency parser is available for Bengali. Development of a full dependency parser is indeed a separate independent research endeavor. To build the Dependency Parser we took part in the ICON 2009 and 2010 NLP TOOLS CONTEST: IL Dependency Parsing. Due to the syntactic richness of Bengali language, hybrid architecture has been proposed for the problem domain. A statistical data driven parsing system (Maltparser) has been used followed by a rule-based post-processing technique. The system has been trained on the ICON NLP TOOLS CONTEST: IL Dependency Parsing datasets. The final system (trained on ICON 2010 NLP TOOLS CONTEST Dataset) has demonstrated an accuracy of unlabeled attachment score (UAS): 81.64%, labeled attachment score (LAS): 54.58%, labeled accuracy score (LS): 50.62% respectively over fine-grained tagset. The details of the parser can be found in (Ghosh et. al., 2009 and Ghosh et. al., 2010) as well as in the appendix section.

2.5.3 Discourse Level Features

Sometime discourse level analysis is very essential for subjectivity detection. Let us have a look at the following English sentence:

```
My camera broke in two days.
```

Detection of subjectivity from the previous sentence is very ambiguous. An incident has been reported but there is no clue as to why the camera broke down. Was the quality of the camera too bad or any accident took place or any other reason was there for the breakage? The issue may be resolved by additional discourse level information from the consecutive sentences. If the following sentence is the successor then probably the previous sentence is not a subjective expression.

```
I felt down from stairs and thus I lost it.
```

On the contrary if the following sentence is the successor in the discourse then undoubtedly the previous sentence expresses subjectivity.

```
I am very upset and decided to buy my next camera, with longer battery
life rather plenty of features.
```

It has been observed during various experiments that every writer generally follows certain pattern during their writing. People start writing (the Letters to the Editor section) by giving their opinion and gradually produce supporting arguments in the successive paragraphs. It has been identified that the following discourse features are useful for subjectivity detection.

2.5.3.1 Positional Aspect

Depending upon the position of subjectivity clue, every document is divided into a number of zones. Various values of this feature are Title of the document, the first paragraph and the last two sentences. A detailed study was carried out on the MPQA and the Bengali corpus to identify the roles of the positional aspect (first paragraph, last two sentences) in the sentence level subjectivity detection task. It has been observed that generally first paragraph and last two sentences of any document contain subjectivity. Corpus statistics prove the phenomenon as reported in the Table 2.3. 48.0% and 56.8% first paragraphs carry subjective information in the MPQA and in the Bengali corpus respectively, whereas 64.0% and 78.0% of last two sentences carry subjective information in the MPQA and in the Bengali corpus respectively. Zone wise statistics could not be prepared for the English IMDB corpus and Bengali BLOG corpus because the corpora are not available as a document.

2.5.3.1.1 Document Title

It has been observed that the Title of a document always carries some meaningful subjective information. Thus a Thematic expression or the sentences containing any of the title words (words that are present in the title of the document) always get higher score.

2.5.3.1.2 First Paragraph

People usually give a brief idea of their beliefs and speculations in the first paragraph of the document and subsequently elaborate or support their ideas with relevant reasoning or factual information. This first paragraph information is useful in the detection of subjective sentences with thematic expressions.

2.5.3.1.3 Last Two Sentences

The general writing style is that every document concludes with a summary of the opinions expressed in the document in the last two sentences. Thus, the last two sentences in a document carry subjective information.

Positional Factors	Percentage	
	MPQA	Bengali
First Paragraph	48.00%	56.80%
Last Two Sentences	64.00%	78.00%

Table 2.3: A Corpus Statistics on Document Level Positional Aspect of the Subjective Sentences from MPQA and Bengali Corpus

2.5.3.2 Theme Words

The term **theme** refers to the sentimental/opinionated topics of any document. It should be noted that the theme does not refer to the simple topics of a document. For example, if we apply any standard IR based topic detection module on a document (e.g., document *D* on *"Travel Guide for Vizag"*) then we may get a topic bag-of-words (e.g., *Vizag, Araku, RK Beach, Climate, Dolphin* etc). But

sentimental/opinionated information may not be present for each topic in the document. Therefore, **theme** is defined as those specific topics (e.g., *Climate*) for which any sentimental information is present in the document (e.g., *"Being close to the sea the climate of Visakhapatnam has no appreciable seasonal changes except during the dry months when the rise in temperature is higher than it is during the monsoon period, should be avoided for travel."*).

Highly inspired by (Wiebe, 2000), a rule-based Theme detection technique has been proposed which has been described in section 2.6.1. Finally, the theme of a document is described as a bag-of-words.

2.5.3.3 Term Distribution Model

The distribution model of a word / term has been proposed as an alternative to the classical Term Frequency – Inverse Document Frequency (TF-IDF) weighting mechanism of standard Information Retrieval (IR). The model characterizes and captures the informativeness of a word by measuring how regularly the word is distributed in a document. (Carenini et al, 2006) introduced the opinion distribution function feature to capture the overall opinion distributed in a corpus. The objective is to estimate the distribution pattern of the k occurrences of the word w_i in a document *d*. Zipf's law describes distribution patterns of words in an entire corpus. In contrast, term distribution models capture regularities of word occurrence in subunits of a corpus (e.g., documents, paragraphs or chapters of a book). A good understanding of the distribution patterns is useful to assess the likelihood of occurrences of a word in some specific positions (e.g., first paragraph or last two sentences) of a unit of text. Most term distribution models try to characterize the informativeness of a word identified by inverse document frequency (IDF). In the present work, the distribution pattern of a word within a document formalizes the notion of topic-sentiment informativeness. This is based on the Poisson distribution. Significant Theme words are identified using TF, Positional and Distribution factor. The distribution function for each theme word in a document is evaluated as:

$$f_d(w_i) = \sum_{i=1}^{n} \frac{(S_i - S_{i-1})}{n} + \sum_{i=1}^{n} \frac{TW_i - TW_{i-1}}{n} \quad \text{----- (2.1)}$$

where n=number of sentences in a document with a particular theme word say TW_i, S_i=sentence id of the current sentence containing the theme word TW_i and S_{i-1}=sentence id of the previous sentence containing the same theme word but as it occurs earlier in the document it is marked as TW_{i-1}. TW_i is the positional id of the current Theme word and TW_{i-1} is the positional id of the same Theme word but in a previous position.

Distribution function for thematic words plays a crucial role during the Thematic Expression identification stage. The distance between any two occurrences of a thematic word measures its distribution value. Thematic words that are well distributed throughout the document are important thematic words. In the learning phase, experiments are carried out using the MPQA Subjectivity word list distribution in the corpus and encouraging results have been observed to identify the theme of a

document. These distribution rules are identified after analyzing the English corpus and the same rules are applied to Bengali corpus as well.

2.6 Subjectivity Adaptation – the Computational Approach

Work in sentiment analysis and classification often assumes that the incoming documents are opinionated. Sentiment analysis systems make false hits while attempting to compute the polarity values for non-subjective or factual sentences or documents. The sentiment analysis systems must decide whether a given document contains subjective information or not and must identify which portions of the document are subjective or factual.

A series of experiments have been carried out starting from the rule based theme detection technique to machine learning techniques such as Conditional Random Field (CRF) and Genetic Algorithm. The details about the systems are elaborated below.

2.6.1 Rule based Theme Detection

The rule based theme detection algorithm identifies subjective sentences in text documents. It first captures discourse level opinion theme in terms of thematic expressions which best describe the opinionated theme of a document. In the next level the algorithm examines the presence of an opinionated evaluative expression associated with the thematic expressions in any sentence. The identification of the most concise feature set and effective construction of the rules for the two stage identification problem are the most important tasks. Experiments have been carried out with an initial list of features and finally some of the features are discarded as they are found to have no contribution towards increasing the system performance. The Theme detection technique has been applied on both English and Bengali language texts (Das and Bandyopadhyay, 2009(a));(Das and Bandyopadhyay, 2009(b));(Das and Bandyopadhyay, 2009(c)).

Many supervised and unsupervised techniques have been explored for subjectivity annotation task by various researchers over a long period of time. Several linguistic resources and tools like Dependency Parsing, Named Entity Recognition, Morphological Analyzer, Stemmer, SentiWordNet, and WordNet etc. have been used several times in the subjectivity detection task. But in the case of morphologically rich Indian languages like Bengali, such resources and tools are not readily available. Highly inspired by (Wiebe et.al, 2005) the present work is initiated to develop a subjectivity classifier that will work on un-annotated text documents. The aim is to design an automatic process that learns linguistically rich extraction patterns for subjective expressions and produces a rich ontological language-specific (rather than domain dependent) knowledge.

The present rule based Theme detection technique works in various steps. First, the system identifies a large set of high frequent words from each corpus (i.e., English: NEWS and Movie Review, Bengali: NEWS and BLOG) that belong to either of Noun, Adjective, Adverb or Verb POS categories. It is assumed

that only words with these POS categories can contribute to the theme of a document (Hatzivassiloglou et. al., 2000). Once the initial list of theme words has been generated then the system assigns a numeric weight to each theme word. Finally, the theme words whose numeric weight is greater than the pre-defined threshold value (identified experimentally) are kept and the rest theme words are discarded.

During the weight assignment, the system looks into the characteristics and behavioral details of the theme words. Each theme word is checked for its fine grained POS category and its presence in a domain dependant ontology list. For Nouns, the named entities (POS tag NNP) get higher weight than common nouns (POS tag NN). A domain ontology list has been developed semi-automatically to check whether a word is high frequent due to the domain or it reflects the theme of a particular document. The functional words are automatically removed from the high frequent theme word list. For English the stop word list (637 entries) is collected from Web[11] and the list (approximately 1000 entries) for Bengali is created manually. After stop word removal the updated theme word list is manually checked to prepare the final ontology list for each domain and language. The system further checks for the syntactic and discourse level behavior of the word. The rule based technique relies on the positional aspect to understand the syntactic behavior of any word. English sentences generally follow SVO (Subject-Verb-Object) pattern while Bengali sentences follow SOV pattern. Therefore, the thematic nouns generally occur at the beginning of any sentence for both English and Bengali. The system assigns a weight to each theme word based on its position.

$$tw_i = \sum_{i=0}^{k} \frac{p_k}{n_k} \quad \text{----- (2.2)}$$

where tw_i is the weight assigned to the i-th theme word, p_k is the position of the theme word in the k^{th} sentence and n_k is the total number of words in the sentence. Discourse level analysis checks whether the word is a title word (tw_i*+=1*) or is present within the first paragraph (tw_i*+=0.5*) or in last two sentences (tw_i*+=0.5*). Accordingly, the weight of the theme word is increased. The incremental values are chosen experimentally.

2.6.1.1 Performance

The Baseline systems for both English and Bengali have been developed using the rules that are based on two primary features, i.e., frequency and positional information. The baseline system evaluation results are shown in the Table 2.4. Further incremental improvement of the baseline system depends on the selection of appropriate additional features. During these experiments, some of the features are discarded as they are found to have no contribution towards increasing the system performance. The final list of features for which any incremental improvement towards system performance is observed is reported in the Table 2.5. The graphical representation of the incremental improvement in the system performance is shown in Figure 2.4 for both the languages. It may be observed that the positional feature and the term distribution feature play very crucial roles to identify the sentence subjectivity. The

[11] http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/

evaluation results of the rule based theme detection technique clearly show an improvement over the evaluation results of the baseline systems.

Language	Precision	Recall
English	51.00%	61.26%
Bengali	49.86%	58.66%

Table 2.4: Results on Subjectivity Baseline System

Feature Set
Frequency
Positional Aspect
Average Distribution
Stemming Cluster
Part of Speech
Chunk
Functional Word
Sentiment Words
Ontology List

Table 2.5: Feature Set for Theme Based Subjectivity Detection

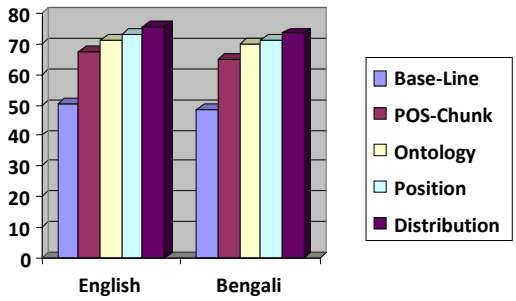


Figure 2.4: Feature Wise Subjectivity Performance by Rule based Theme Detection

2.6.2 Theme Detection through Machine Learning: The CRF based Approach

The accuracy level of the rule based theme detection system is approximately 50%. During error analysis, it has been identified that theme identification and subjectivity detection are deep semantic issues and it is nearly impossible to develop a complete set of definite rules. To overcome the limitations of the rule based system a machine learning module has been developed with the already identified features along with a few additional ones. The Conditional Random Field (CRF)[12] machine learning algorithm has been used. The CRF base subjectivity detection has achieved precision values of 76.08% and 79.90% for English NEWS and movie review corpus and 72.16% and 74.6% for Bengali news and blog domains respectively (Das and Bandyopadhyay, 2009(a)).

Some additional features have been included during the training of the CRF based classifier. These features include the dependency relations and named entities. Dependency relations are extracted using the Stanford dependency parser[13] for English and the Bengali dependency parser (Ghosh et. al., 2009). The Bengali dependency parser has been developed as a joint development work in the laboratory during the present work and has been discussed in detail in the Appendix section. English named entities are identified using the Stanford NER[14]. For Bengali named entity recognition the NER system (Ekbal et. al., 2008(a)) has been used.

2.6.2.1 Performance

The effectiveness of each feature, in terms of increment in the precision score of the system, in the CRF-based Subjectivity detection tasks for English and Bengali are presented in Table 2.6. The precision and recall values of the subjectivity classifier are shown in Table 2.7 for the various English and Bengali corpora. It may be observed that subjectivity detection is trivial for review corpus and blog corpus rather than for news corpus. There is more factual information in the news corpus than in the review or blog corpus that generally contain people opinion.

Feature Ablation	Overall Performance Incremented By	
	English	Bengali
Stemming Cluster	5.32%	4.05%
Part Of Speech	4.12%	3.62%
Chunk	3.98%	4.07%
Average Distribution	2.53%	1.88%
Sentiment Lexicon	6.07%	5.02%
Positional Aspect	3.06%	3.66%

Table 2.6: Feature Wise Subjectivity Performance Improvement

[12] http://crfpp.sourceforge.net
[13] http://nlp.stanford.edu/software/lex-parser.shtml
[14] http://nlp.stanford.edu/ner/index.shtml

Languages	Domain	Precision	Recall
English	MPQA	76.08%	83.33%
	IMDB	79.90%	86.55%
Bengali	NEWS	72.16%	76.00%
	BLOG	74.6%	80.4%

Table 2.7: The Overall Subjectivity Performance for English and Bengali using CRF

2.6.3 Adaptive Genetic Algorithm: Multiple Objective Optimization

Machine learning algorithms in NLP generally experiment with combination of various syntactic and semantic linguistic features to identify the most effective feature set. The sentiment detection task has been viewed this as a multi-objective or multi-criteria optimization search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set through the process of crossover and mutation after each generation.

In the present task, the Genetics Based Machine Learning (GBML) has been used to automatically identify the best feature set based on the principle of natural selection and survival of the fittest. The identified fittest feature set is then locally optimized. Global optimization is achieved by multi-objective optimization technique. The local optimization identifies the best range of feature values of a particular feature. The global optimization technique identifies the best range of values for the multiple features. The proposed technique has been tested for English and Bengali corpus and for the news, movie review and blog domains. The system evaluation results show precision of 90.22%, and 93.00% respectively for English NEWS and Movie Review corpus and 87.65% and 90.6% for Bengali NEWS and Blog corpus (Das and Bandyopadhyay, 2010(g)).

2.6.3.2 Why Genetic Algorithm

Genetic Algorithms (GAs) are probabilistic search methods (Holland, 1975; Goldberg, 1989). GAs are applied for natural selection and natural genetics in artificial intelligence to find the globally optimal solution from the set of feasible solutions. Nowadays GAs are being applied in various domains that include timetable, scheduling, robot control, signature verification, image processing, packing, routing, pipeline control systems, machine learning, and information retrieval (Kraft, 1994; Bautista et. al, 1997).

There is only one effort that attempted Genetic Algorithm (Abbasi et. al, 2008) for the opinion mining problem. They developed the Entropy Weighted Genetic Algorithm (EWGA) for opinion feature selection. The features and techniques result in the creation of a sentiment analysis approach geared towards classification of web discourse sentiments in multiple languages. The EWGA has been applied for English and Arabic languages. It uses the information gain (IG) heuristic to weigh the various opinion attributes. They have compared their results with a SVM based method and with previous existing methods in the literature. Table 2.8 report the features used for the GA based subjectivity detection for

English and Arabic. Table 2.9 shows the taxonomies for sentiment/subjectivity detection. Table 2.10 shows selected previous studies dealing with sentiment/subjectivity detection based on the proposed taxonomy. The EGWA method outperforms the existing subjectivity classification methods and has achieved approximately 94.00% accuracy score on both the English and Arabic languages. There is a clear resemblance with the feature selection in the present work as reported in the Table 2.2. It is very clear from the Table 2.10 that no GA based subjectivity detection method exists in the literature. Only (Abbasi et. al, 2008) has attempted with the GA based method for English and Arabic. In the present work, the GA based subjectivity detection method has been attempted for English and Bengali. In both the cases the accuracy figures are relatively higher than all the previous proposed methods. It clearly establishes the efficiency of GA mechanism for sentiment/subjectivity detection.

Category	Feature Group	Examples
Syntactic	**POS N-grams**	frequency of part-of-speech tags (e.g., NP_VB)
	Word Roots	varies frequency of roots (e.g., slm, ktb)
	Word N-Grams	varies word n-grams (e.g. senior editor, editor in chief)
	Punctuation	occurrence of punctuation marks (e.g., !;:,.?)
Stylistic	**Letter N-Grams**	frequency of letters (e.g., a, b, c)
	Character N-Grams	varies character n-grams (e.g., abo, out, ut, ab)
	Word Lexical	total words, % char. per word
	Character Lexical	total char., % char. per message
	Word Length	frequency distribution of 1-20 letter words
	Vocabulary Richness	richness (e.g., hapax legomena, Yule's K)
	Special Characters	occurrence of special char. (e.g., @#$%^&*+)
	Digit N-Grams	varies frequency of digits (e.g., 100, 17, 5)
	Structural	has greeting, has url, requoted content, etc.
	Function Words	frequency of function words (e.g., of, for, to)

Table 2.8: English and Arabic Feature Sets (Abbasi et. al, 2008)

Category	Example	Label
Features		
Syntactic	Word/POS tag n-grams, phrase patterns, punctuation	F1
Semantic	Polarity tags, appraisal groups, semantic orientation	F2
Link Based	Web links, send/reply patterns, and document citations	F3
Stylistic	Lexical and structural measures of style	F4
Techniques		
Machine Learning	Techniques such as SVM, Naïve Bayes, etc.	T1
Link Analysis	Citation analysis and message send/reply patterns	T2
Similarity Score	Phrase pattern matching, frequency counts, etc.	T3
Domains		
Reviews	Product, movie, and music reviews	D1
Web Discourse	Web forums and blogs	D2
News Articles	Online news articles	D3

Table 2.9: Taxonomy of Sentiment / Subjectivity Detection (Abbasi et. al, 2008)

Study	Features				Reduced Feat.	Techniques			Domains			No. Lang.
	F 1	F 2	F 3	F 4	Yes/No	T 1	T 2	T 3	D 1	D 2	D 3	1-n
Subasic & Huettner, 2001	••	••			No			••			••	1
Tong, 2001	••	••			No			••	••			1
Morinaga et al., 2002	••				Yes			••	••			1
Pang et al., 2002	••				No	••			••			1
Turney, 2002	••	••			No			••				1
Agrawal et al., 2002	••		••		No	••	••			••		1
Dave et al., 2003	••				No	••		••	••			1
Nasukawa & Yi, 2003	••	••			No			••	••			1
Riloff et al., 2003		••		••	No	••					••	1
Yi et al., 2003	••	••			Yes			••	••		••	1
Yu & Hatzivassiloglou,	••	••			No	••		••			••	1
Beineke et al., 2004		••			No	••		••	••			1
Efron, 2004	••		••		No	••	••			••		1
Fei et al., 2004		••			No			••	••			1
Gamon, 2005	••			••	Yes	••			••			1
Grefenstette et al., 2004	••	••			No			••		••		1
Hu & Liu, 2004	••	••			No			••	••			1
Kanayama et al., 2004	••	••			No			••	••			1
Kim & Hovy, 2004		••			No			••		••		1
Pang & Lee, 2004	••	••			No	••		••	••			1
Mullen & Collier, 2004	••	••			No	••			••			1
Nigam & Hurst, 2004	••	••			No	••				••		1
Wiebe et al., 2005	••			••	Yes	••		••		••	••	1
Liu et al., 2005	••	••			No			••	••			1
Mishne, 2006	••	••		••	No	••				••		1
Whitelaw et al. 2005	••	••			No	••			••			1
Wilson et al., 2005					No	••					••	1

Table 2.10: Selected Previous Studies in Sentiment Polarity Classification (Abbasi et. al, 2008)

2.6.3.3 Basic Principles of Genetic Algorithm

GAs are characterized by the five basic components. Figure 2.5 displays a diagrammatic representation of the whole process. The basic system components are:

1. Chromosome representation for the feasible solutions to the optimization problem.

2. Initial population of the feasible solutions.
3. A fitness function that evaluates each solution.
4. Genetic operators that generate a new population from the existing population.
5. Control parameters such as population size, probability of genetic operators, number of generation etc.

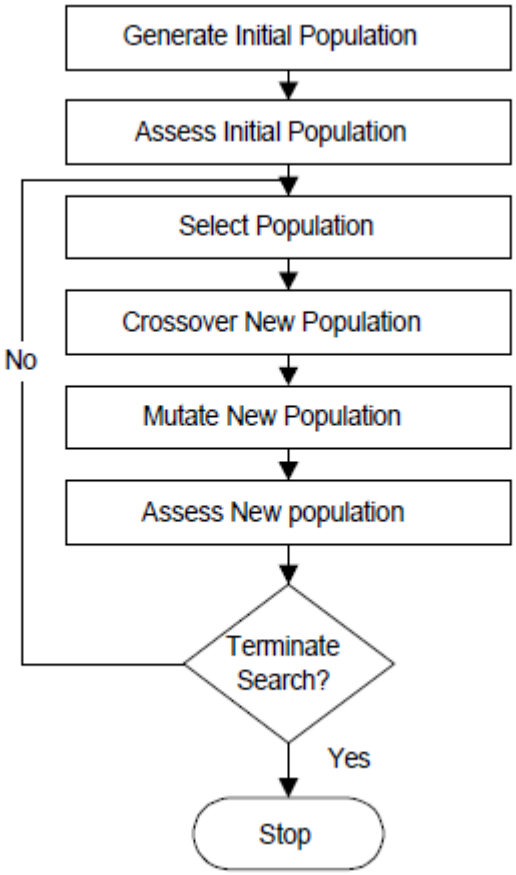


Figure 2.5: The Overall Process of Genetic Algorithm

2.6.3.4 Proposed Technique

The experiment starts with a large set of possible extractable set of syntactic, semantic and discourse level features. The fitness function calculates the accuracy of the subjectivity classifier based on the fittest feature set identified by natural selection through the process of crossover and mutation after each generation. Sometimes, problems need to be formulated with more than one objective, since a single objective with several constraints may not adequately represent the problem being faced. If so, there is a vector of objectives/features that must be traded off in some way. The relative importance of these objectives/features is not generally known until the system's best capabilities are determined and

tradeoffs between the objectives fully understood. As the number of objectives/features increases, tradeoffs are likely to become complex and less easily quantified. Thus, requirements for a multi-objective design strategy must enable a natural problem formulation to be expressed, and be able to solve the problem and enter preferences into a numerically tractable and realistic design problem. Genetic algorithms have been used for this kind of natural problem formulation. The subjectivity classification problem can be viewed as a summation of the subjectivity probabilities of the set of possible features as shown in the following equation, with the hypothesis of Multi-objective Optimization. Multi-objective optimization is concerned with the minimization of a vector of objectives $F_i(x)$ that can be the subject of a number of constraints or features.

$$f_s = \sum_{i=0}^{N} F_i(x) \quad \text{----- (2.3)}$$

where f_s is the resultant subjectivity function to be calculated and $F_i(x)$ is the i^{th} feature function. If the present model is represented in a vector space then the above function can be rewritten as:

$$f_s = [F_1(x), F_2(x),...,F_m(x)] \quad \text{----- (2.4)}$$

This equation specifies what is known as the dot product between vectors. Now, in general, the dot product between two vectors is not particularly useful as a classification metric, since it is too sensitive to the absolute magnitudes of the various dimensions. From the previous research it is already proven that particular features like Syntactic Chunk Label and Discourse Level feature have their own range of tentative values. For example, words of some specific POS category reflect sentiment very well, hence it can be inferred that frequent occurrence of these POS category words in a sentence increases the subjectivity value of the sentence. Further, occurrence of low frequency words is a well established clue of subjectivity but a sentence with only low-frequent words may not be subjective always. In a multiple feature or multiple vector space models the desired optimal solution may be obtained by finding out the optimal range of every feature vector. Hence it is obvious that in single-criterion optimization, the notion of optimality scarcely needs any explanation in this particular category of problem. We simply seek the best value of assumedly well-defined multi-objective (utility or cost) optimization function.

2.6.3.5 Problem Formulation

To maximize the subjectivity probability, the occurrence of low-frequency words (LFW), title words (TW), average distributed words (ADW) (obtained by term distribution model) and theme words (TD) and their position in each sentence are calculated. The matrix representation for each sentence looks like:

[X, Y]= [frequency in the entire corpus, position in the sentence]

Example Sentence: *Weiner's district has a substantial Jewish^{LFW} population, about one third of the electorate^{ADW}, so there is credible speculation that President Obama's^{TW} low approval ratings on the issue of Israel are responsible for his unpopularity in New_York^{TD} 9.*

Therefore, the matrix representation for the above sentence will be:

LFW= [5, 6]: for the word "*Jewish*"

TW= [34, 9]: for the word "*Obama*"

ADW= [13, 10]: for the word "*electorate*": stem cluster form: "*election, electorate,..*"

TD= [36, 15]: for the word "*New York*"

The above data are plotted as frequency (X-axis) versus position (Y-axis) in the Figure 2.6.

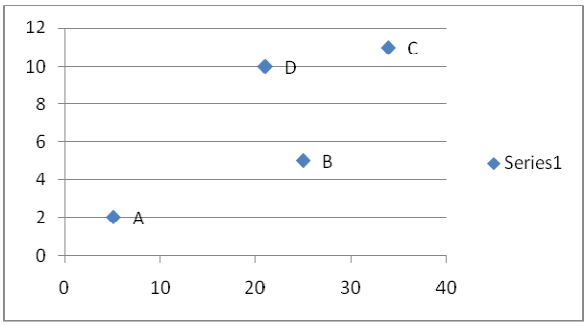


Figure 2.6: Frequency vs. Position Plot of Subjective Words

Note that because f_s is a vector, if any of the components of $F_i(x)$ are competing, there is no unique solution to this problem. Instead, the concept of Pareto optimality must be used to characterize the objectives. To define this concept more precisely, a feasible region may be considered in a Pareto-optimal plane. Scanning the graph reveals that the best points are lower and to the right of the plot. In particular, scenarios A, B and C seem like good possible choices: even though none of the three points is best along both dimensions, there is trade-offs from one of these three scenarios to another; there is gain along one dimension and loss along the other. In optimization terminology these three points are non-dominated because there are no points better than these on all criteria.

The GBML provides the facility to search in the Pareto-optimal set of possible features. This Pareto-optimal set is being generated from crossover and mutation. To make the Pareto optimality mathematically more rigorous, it can be stated that a feature vector x is partially less than feature vector y, symbolically x<p y, when the following condition holds:

$$(x < p\ y) \Leftrightarrow (\forall_i)(x_i \le y_i) \wedge (\exists i)(x_i < y_i) \quad \text{----- (2.5)}$$

This may be mapped to Pareto plane as shown in Figure 2.7, where Pareto front of non-dominated points are highlighted in red color.

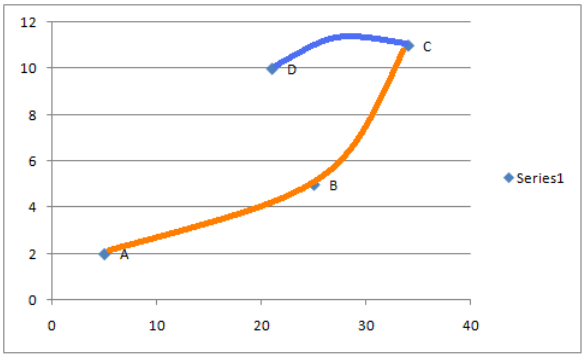


Figure 2.7: Pareto Plane of Frequency vs. Position Plot

In the notion of Pareto optimality by multi-objective optimization, the GA has been used in parallel fashion. The methodology used is as follows:

1. Generate chromosome for each feature.
2. Initialize population for each feature.
3. For i=1 to population size
4. For j=1 to feature vector size
5. Compute fitness P (t). (where t is the current generation)
6. If termination condition is satisfied go to Step 10.
7. Select (P)
8. Crossover (P)
9. Mutate (P)
10. Go to Step 3.
11. Output
12. End

The parallelism is obtained by generating n number of GA based subjectivity classifiers. Based on the principle of survival of the fittest, a few of the feature strings are selected. This parallelism provides the granularity for every feature. The n numbers of GA based subjectivity classifiers are synchronous in nature and they simultaneously generate their population. The fitness value is calculated afterevery iteration. The optimal solution is selected based on the theory of Pareto optimality which helps to reach

the fittest global solution from the local best solution for each feature. The effectiveness of the present technique is observed in the experimental results.

2.6.3.6 Chromosome Representation

The size of the chromosome for every feature varies according to the possible solution vector size. Tentative solutions are made of sequences of genes. Each gene corresponds to word sequence in the sentence to be tagged.

The chromosomes forming the initial population are created by random selection of one of the valid tags for each word from a dictionary. For the present task, real encoding has been used. A sentence wise feature vector can be represented as,

```
Imperialism/NNP is/VBZ the/DT source/NN of/IN war/NN and/CC the/DT
disturber/NN of/IN peace/NN.
```

The encoded chromosome is represented in Figure 2.8. The real values are the serial number of the corresponding tag from the POS Tag labeled dictionary. Table 2.11 reports how real values vary for every feature.

Features	Real Values
POS	1-21 (Bengali)/1-45 (English)
SentiWordNet	-1 to +1
Frequency	0 or 1
Stemming	1 to 17176/ 1 to 1235
Chunk Label	1-11 (Bengali) / 1-21 (English)
Dependency Parsing	1-30 (Bengali) / 1-55 (English)
Title of the	Varies document wise
First Paragraph	Varies document wise
Average Distribution	Varies document wise
Theme Word	Varies document wise

Table 2.11: Dimension of Chromosome Encoding with Chosen Subjectivity Features

NNP	VBZ	DT	NN	IN	NN	CC	DT	NN	IN	NN
1	12	6	2	18	2	4	6	2	18	2

Figure 2.8: Chromosome Representation for GA Based Subjectivity Detection

The POS feature values vary for languages as the tag sets are different. There are 21 tags and 45 tags in the POS tagset for Bengali and English respectively. The Chunk label and the Dependency relations follow the same mechanism for encoding as the POS feature. Sentiment words from SentiWordNet get the feature value as: -1 for negative, 0 for neutral and +1 for positive words. Only low frequency words are considered to be essential. Any word occurring less than 5 times in the corpus has been considered

as a low frequency word. This feature is encoded as a binary feature: 1 for frequency less than equal to 5 or 0 where frequency is greater than 5. Stems from the corpus are listed and the serial number of any stem within the list is used to encode the chromosome. It is basically the set of unique wordforms in any corpus.

2.6.3.7 Crossover

Crossover is the genetic operator that mixes two chromosomes together to form a new offspring. Crossover occurs only with some crossover probability. Chromosomes that are not subjected to crossover remain unmodified. The intuition behind crossover is the exploration of new solutions and exploitation of old solutions. GAs constructs a better solution by mixing the good characteristics of chromosomes together.

2.6.3.8 Mutation

Mutation involves the modification of the values of each gene of a solution with some mutation probability. The process of mutation changes some values of chromosomes generating the different breeds. Mutant Chromosomes may be better or poorer than old chromosomes. If they are poorer than old chromosomes, they are eliminated in the selection step. The objective of mutation is to restore the lost feature and explore new ones in order to reach the fittest solution. For example, in the following chromosome a random mutation occurs at position 10.

Result: 1 0 1 1 1 1 1 1 0 1 1 1 1 0 1

1 0 1 1 1 1 1 1 0 0 1 1 1 0 1

2.6.3.9 Natural Selection

After the population fitness has been evaluated, the next step is chromosome selection. Selection embodies the principle of 'survival of the fittest'. The mutant fittest chromosomes are selected for reproduction. A few poor chromosomes or lower fitness chromosomes may be selected.

2.6.3.10 Fitness Evaluation

Fitness function is a performance measure or a reward function which evaluates how good each solution is. The following cost-to-fitness transformation is commonly used with GAs.

$$f(x) = C_{\max} - g(x) \text{ when } g(x) < C_{\max} \quad ----- \quad (2.6)$$

$$=0 \qquad \text{otherwise}$$

There are varieties of ways to choose the coefficient C_{max}. C_{max} may be considered as an input coefficient or the largest g value observed thus far or the largest g value in the current population or the largest of the last k generation.

When the natural objective function formulation is a positive utility function there is no difficulty with the direction of the function: maximized desired profit or utility leads to desired performance. But still there are some problems with negative utility function during the fitness evaluation of n number of features. To overcome this, the fitness function is transformed by the following the equation:

$$f(x) = u(x) + C_{\min} \; When \; u(x) + C_{\min} > 0 \quad ----- \quad (2.7)$$

$$= 0 \qquad otherwise$$

For the present problem there is a single fitness function to select the best Pareto optimal plane.

2.6.3.11 Performance

The Java API for Genetic Algorithm[15] application has been used. Approximately 70% of every corpus has been used for training purpose and the rest 30% has been used for testing purpose. The following parameter values are used for the genetic algorithm: population size=50, number of generation=50. The mutation and crossover probabilities are selected adaptively.

Languages	Domain	Precision	Recall
English	MPQA	90.22%	96.01%
	IMDB	93.00%	98.55%
Bengali	NEWS	87.65%	89.06%
	BLOG	90.6%	92.40%

Table 2.12: Results of Final GA based Subjectivity Classifier

The precision and recall values of the subjectivity classifier are shown in Table 2.12 for all the corpora selected for English and Bengali. It is observed that subjectivity detection is trivial for review corpus and blog corpus rather than for news corpus. In news corpus there is more factual information than review or blog corpus that generally contain people's opinion. Thus subjectivity classification task is domain dependent. But the proposed technique is domain adaptable through the use of natural selection. The difference of GA-based classifier with other statistical systems is that a whole sentence can be encoded in GA and can be used as a feature. In other classifier systems, n-gram method has been followed. The fixed size of n in the n-gram does not fit into the variable string length of an input string.

[15] http://www.jaga.org/

Publications

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010. ***Subjectivity Detection using Genetic Algorithm***. *In the Proceeding of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA10), ECAI 2010*, Pages 14-21, Lisbon, Portugal.
   http://www.amitavadas.com/Pub/GA.pdf

2. **Amitava Das** and Sivaji Bandyopadhyay. 2009. ***Subjectivity Detection in English and Bengali: A CRF-based Approach***. *In the Proceeding of the International Conference on Natural Language Processing (ICON 2009)*, Pages 358-363, Hyderabad, India.
   www.amitavadas.com/Pub/ICON_Final_Amitava.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2009. ***Theme Detection an Exploration of Opinion Subjectivity.*** *In the Proceeding of the Affective Computing & Intelligent Interaction (ACII2009)*, Pages 1-6, Amsterdam, Netherlands.
   http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5349599

4. **Amitava Das** and Sivaji Bandyopadhyay. 2009. ***Extracting Opinion Statements from Bengali Text Documents through Theme Detection.*** *In the Proceeding of the 17th International Conference on Computing (CIC-09), GEOS 2009*, Mexico City, Mexico.
   www.amitavadas.com/Pub/CIC_Poster_2008_JU.pdf

Chapter 3

Sentiment / Opinion Polarity Detection

The ***polarity classification*** is the classic problem from where the cultivation of Sentiment Analysis (SA) has actually started. The problem of ***polarity classification*** involves sentiment/opinion classification into semantic classes such as *positive, negative or neutral* and/or other fine-grained emotion classes like *happy, sad, anger, disgust and surprise*. One of the most noteworthy earliest research works on sentiment polarity classification has been conducted by (Turney et. al., 2002) with review corpus. The semantic classes were considered as "*thumbs up*" or "*thumbs down*" for movie reviews. Motivated by different real-world applications, researchers have considered a wide range of semantic classes over a variety of different types of corpora or problem domains. The development of a fully automatic polarity classifier is still the basic requirement to meet the real life needs and the ultimate desire of the whole Sentiment Analysis research.

In this chapter we will describe about the various polarity classification techniques, proposed by us. The chapter is organized as follows. There are several factors that make the automatic polarity classification a very challenging research problem. These factors are discussed in section 3.1. The various research attempts by several researchers who attempted to formulate the research problem and the solution for the sentiment polarity classification have been discussed in section 3.2. The polarity classification technique has been developed for the Bengali language which is a resource poor language. Thus, acquisition of relevant resources and the development of appropriate tools is one of the important aspects of the present work. The resource acquisition process includes corpus collection and annotations. A dependency parser for Bengali has also been developed which is a necessary tool to detect syntactic sentimental semantics from text. The acquisition of relevant resources and the development of appropriate tools have been discussed in the section 3.3. The details of the syntactic polarity classification technique are described in the section 3.4. During the error analysis of the syntactic polarity classifier, it has been observed that the performance of the polarity classifier mainly drops for the unknown or new words. These observations are described in detail in the section 3.5. A closer look at the error analysis points to questioning the two standard steps in the polarity classification class, the use of a prior polarity lexicon followed by the application of any NLP technique. A new method has been proposed which uses a lexical network based on Vector Space Model (VSM) that holds the contextual sentimental polarity. The problem of holding sentiment knowledge with context is defined as Sentimantics in the present work. The motivation for the proposed Sentimantics and its fundamentals has been discussed in the section 3.6. The details of the proposed lexical networks to solve the Sentimantics are described in the section 3.7.

3.1 Understanding Sentiment: The Social Norms

The sentiment analysis research started as a content analysis research problem in the behavioral science. The General Inquirer System (1966) (Stone, 1966) is the first attempt in this direction. The aim was to gain understanding of the psychological forces and the perceived demands of the situation that were in effect when the document was written. Although the sentiment analysis research has started

long back, still the question "***What is sentiment or opinion?***" remain unanswered! Moreover no complete set of psychological forces could be defined that really affect the sentiment of the writer.

The sentiment analysis is hard for at least a few reasons. In general, the sentimental knowledge of a human grows with time and social interactions. For example, the word "*selfish*" is a negative word. This knowledge is psychologically disposed in the human brain. To solve the polarity identification problem researchers trust on prior polarity lexicons. A prior polarity lexicon is a set of sentiment words associated with prior polarity scores. But the main difficulty with the prior polarity lexicon is that it is very hard to linguistically determine the prior polarity of any lexicon because the prior polarity is not a linguistic property, rather it is the cognitive property of human intelligence gathered by social interactions.

Therefore, beyond the issues of ambiguity, it is hard for computers to pull out the meaning in a statement or a set of statements because people express things in different ways. Finding the sentiment in a sentence is very difficult using some automatic techniques.

3.2 Previous Studies

The development of the General Inquirer System[1] (1966) (Stone, 1966) by Philip Jones and colleagues in Harvard was probably the first milestone to identify textual sentiment. They called it a content analysis research problem in the field of behavioral science. The aim was to gain understanding of the psychological forces and the perceived demands of the situation that were in effect when the document was written. The system usually counts the positive or negative emotion instances. General Inquirer is empowered with manually classified terms labeled with various types of positive or negative semantic orientations, and the words in the input have to reach an agreement or disagreement with the stored list. The lexicon used in the Inquirer system has been classified into several categories such as positive, negative, pleasure, feel, need, goal, place, vehicle etc. The rich lexical resource of General Inquirer[2] has been further used by various researchers to develop their sentiment or affect lexicon. Automatic identification of semantic orientation of a text is the initiation of today's sentiment polarity classification problem.

Two types of identification problems have been addressed by previous researchers, Sentimental polarity identification ("*The text is positive or negative?*") and magnitude identification of any sentimental text ("how positive or negative is it?"). The previous research endeavors have been discussed in the following three subsections.

[1] http://www.wjh.harvard.edu/~inquirer/
[2] http://www.wjh.harvard.edu/~inquirer/homecat.htm

Section 3.2.1 will cover the fundamental background of polarity classification problem. Specifically, the section will discuss the key concepts of the common formulations of classification problems in sentiment analysis and opinion mining by using ***prior polarity*** lexicon.

Section 3.2.2 is devoted to an in-depth discussion of the different types of approaches to classification, regression and ranking problems. Various researchers realized that only the prior polarity method may not excel alone and NLP or other techniques are required for contextual polarity disambiguation.

Section 3.2.3 will discuss about the very recent trends in Sentiment Analysis. These techniques take a different way for sentiment knowledge representation. It follows the mental lexicon model to hold the contextual polarity like human psychological knowledge representation.

3.2.1 Prior Polarity Lexicon

The polarity classification problem started as a semantic orientation determination problem. Peter Turney and Vasileios Hatzivassiloglou are the pioneers who started the initial experimentations during early 90's. In the year of 1997, Hatzivassiloglou identified the semantic orientation of adjectives. This is the first research attempt that provided the effective and empirical method of building sentiment lexicon. After a few years, Peter Turney came up with his revolutionary approach *Thumbs Up* and *Thumbs Down* for positive and negative review classification. Finally, the concept of prior polarity lexicon evolved and firmly established itself with the innovation of SentiWordNet by Andera Esuli in 2004. All the present polarity classifiers follow a two step methodology. In the first step, classifiers identify the polarity of a text by using any dictionary of prior polarity lexicon and in the next step contextual polarity is disambiguated with the help of NLP techniques or any other fine-grained techniques. In this section, the fundamental works are mentioned that have established the theory of prior polarity lexicons. Prior polarity lexicon involves semantic orientation determination from a text and is a big challenging research issue itself. The various semantic orientation determination techniques suggested by various previous researchers are now discussed.

(Hatzivassiloglou et. al., 1997) proposed their log-linear regression model to predict the orientation of conjoined adjectives. The log-linear regression model uses the number of constraints identified from large corpus and clusters the conjoined adjectives into finite number of groups of different orientations which are finally labeled as positive or negative. The approach relies on some linguistic features, or indicators, with semantic orientation of conjoined adjectives, syntactically co-occurred. They followed the hypothesis that the conjoined adjectives usually are of the same orientation, for example, ***fair*** and ***legitimate***, ***corrupt*** and ***brutal***. The system is trained on a large corpus to identify these relations to predict the semantic orientation of the conjoined adjectives that are linguistically anomalous. The situation is reversed for "***but***", which usually connects two adjectives of different orientations, for example, ***short*** but ***good***, ***far*** but ***comfortable***. The system identifies and uses this indirect information in the following stages:

1. All conjunctions of adjectives are extracted from the corpus along with relevant morphological relations.

2. A log-linear regression model combines information from different conjunctions to determine if each of the conjoined adjectives is of same or different orientation. The result is a graph with hypothesized same- or different-orientation links between adjectives.

3. A clustering algorithm separates the adjectives into two subsets of different orientation. It places the words of same orientation into the same subset.

The average frequencies in each group are compared and the group with the higher frequency is labeled as positive.

This is one of the most important milestones for textual sentiment analysis research. The performance of the reported system is quite high. But the research endeavor is also important for other important aspects such as problem definition and formulation of several hypotheses that needs to be checked further for validity.

- The requirement of an automatic system for detecting the non-linguistic characteristics like semantic orientation of text is established although the authors have suggested the system for adjectives only.

- Syntactically co-occurred adjectives belong to the same semantic orientation group although there are some exceptional cases for the conjunction "*but*" and others.

The problem definition has motivated other researchers to pursue the research problem. One of the most cited research papers in the literature is written by (Turney, 2002). Turney devised an algorithm to extract Pointwise Mutual Information (PMI) for consecutive words and their semantic orientation. The experiments have been carried out on movie review corpus and thus the author referred the semantic orientations as "***thumbs up***" or "***thumbs down***" instead of positive or negative (Hatzivassiloglou et. al., 1997). The simple syntactic patterns considered for the experiments have been described in Table 3.1.

First Word	Second Word	Third Word(Not Extracted)
JJ	NN or NNS	Anything
RB, RBR, or RBS	JJ	not NN nor NNS
JJ	JJ	not NN nor NNS
NN or NNS	JJ	not NN nor NNS
RB, RBR, or RBS	VB, VBD,VBN, or VBG	Anything

Table 3.1: Syntactic Patterns of POS tags for Pointwise Mutual Information (PMI) Calculation (Turney, 2002)

The Brill POS Tagger (Brill, 1994)[3] tagger has been used for the task. Phrases containing words with adjective, adverb, noun and verb words have been extracted as such words depict diverse semantic information. After such phrases are extracted the PMI algorithm executes a Latent Semantic Analysis on these phrases to determine their semantic orientation. During the initial phases of sentiment analysis research people generally believed in syntactic influence on the semantic orientation of words. To investigate these relationships in real corpora they generally started with hand-crafted lexicons. Turney used only 1336 hand-labeled adjectives as the seed words.

Most of the semantic orientation detection tasks started with binary classifications, e.g., positive/negative, thums up/thums down, pro/con, like/dislike etc. But gradually a group of independent researchers started thinking about more fine-grained classifications for semantic orientations. They named it emotion analysis or affect computing with the wide perception that the future of human-computer interaction lies in themes such as entertainment, emotions, aesthetic pleasure, motivation, attention, engagement, etc. One of the most important research endeavors in this genre is by (Valitutt et al., 2004). The authors developed a preliminary version of a lexical knowledge base containing words in an affective lexicon connected with a set of affective concepts. This resource (named WORDNET-AFFECT) was developed starting from the lexical knowledge base WORDNET, through a selection and labeling of the affective concepts (represented by sets of synonyms). WORDNET-AFFECT was then extended taking into account OpenMind, a database of common sense sentences, in which there is a considerable amount of common sense knowledge (Singh, 2002). WORDNET-AFFECT is also a prior polarity lexicon resource but the semantic classes used here are n-nary such as anger, doubt, competitive, skepticism and pleasure etc.

In the year of 2006, Esuli and Sebastiani (Esuli and Sebastiani, 2006) introduced the idea of SentiWordNet[4] which became the most widely used lexical resource for sentiment analysis in the successive years. It is a semi-automatically developed lexical resource, which holds WordNet synsets and prior polarity scores as positivity and negativity. The total occurrence of a particular word in a domain corpus is counted as well as its positive and negative occurrences. Let us consider that the total occurrence of the word "*long*" in a domain corpus is *n* and the positive and negative occurrences of the word are S_p and S_n respectively. Therefore in the developed sentiment lexicon the assigned positivity and negativity scores of the word will be calculated as follows:

$$\text{Positivity} = \frac{S_p}{n}$$

$$\text{Negativity} = \frac{S_n}{n}$$

Four years later, in 2010, the authors released the next version of the resource called SentiWordNet 3.0.

[3] http://www.ling.gu.se/~lager/mogul/brill-tagger/index.html
[4] http://sentiwordnet.isti.cnr.it/

(Mihalcea et. al., 2007) have proposed a nice architecture for the development of subjectivity lexicon for resource scarced Romanian language. They started with a small set of seed words for four POS categories: noun, verb, adverb and adjective. The initial seed word list is incremented with an online dictionary along with a small set of manually annotated corpora in a bootstrapping manner.

Subjectivity lexicon (Wilson et. al., 2005) is one of the widely used English sentiment lexicon mainly developed from news corpora. The authors showed that lexico-syntactic patterns such as:

X-Drive

Y-got-Angry

help to identify subjective expressions across domains. A subjectivity classifier has been trained on a manually annotated data set and has been used to annotate more data. The data is then used to train the system again by bootstrapping method.

(Denecke, 2009) provides an interesting study with the prior polarity scores from the SentiWordNet and shows how these scores could be useful for multiple domains. Two methodologies, one rule-based and another machine learning based, have been proposed in the work. The positivity, negativity and the objectivity scores have been used from the SentiWordNet. A noticeable accuracy has been achieved with the machine learning approach.

(Ohana and Tierney, 2009) have reported their experimentation on review classification using SentiWordNet, which proves the credibility and acceptability of this kind of lexicon resources. A method has been proposed for applying SentiWordNet to derive a data set of document metrics and other relevant features. Experiments have been performed on sentiment classification of film reviews using the SentiWordNet polarity data set.

Besides the semantic orientation detection techniques, a number of researchers have attempted for sentiment strength detection. (Thelwall et. al., 2010) have proposed methods for the sentiment strength detection from short informal text. In addition to the research effort concerning the strength detection for multiple emotions (Strapparava and Mihalcea, 2008), there are some works on positive-negative sentiment strength detection. One previous study has used modified sentiment analysis techniques to predict the strength of human ratings on a scale of 1 to 5 for movie reviews (Pang & Lee, 2005). This is a kind of sentiment strength evaluation with a combined scale for positive and negative sentiment. Sentiment strength classification has also been developed for a three level scheme (low, medium, and high or extreme) for subjective sentences or clauses in newswire texts using a linguistic analysis technique that converts sentences into dependency trees reflecting their structure (Wilson et al., 2006).

Sentiment analysis researchers have established that prior polarity lexicons are necessary for polarity classification task. Therefore, prior polarity lexicon development endeavor have been noticed for other languages as well, e.g., Chinese (He et. al., 2010), Japanese (Torii et. al., 2010) and Thai (Haruechaiyasak et. al., 2010).

3.2.2 Different Classification Strategies

It has been reported by several researchers that higher accuracy for prior polarity identification is very hard to achieve. Prior polarity values are approximates. Researchers have argued that prior polarity method along with NLP or other techniques are required for contextual polarity disambiguation.

The use of NLP methods or machine learning techniques over the human developed prior polarity lexicon was first pioneered by (Pang et al., 2002). The authors have considered the problem of classifying documents not by topic but by overall sentiment, e.g., determining whether a review is thumbs up (positive) or thumbs down (negative). Using movie reviews data, it has been observed that standard machine learning techniques definitively outperform the human developed prior polarity baseline. However, the three machine learning methods they employed (Naive Bayes, Maximum Entropy classification, and Support Vector Machines) do not perform as well on sentiment classification compared to their performance on traditional topic based categorization. Thereafter a numbers of research attempts like (Salvetti et. al., 2004) have been identified that follow the same system architecture for various other languages and domains.

Another important research attempt to overcome the limitations of the manually augmented prior polarity lexicon is found in (Liu et al., 2003) but the problem domain differs from that of (Pang et al., 2002). Several methods have been presented for assessing the affective qualities, i.e., emotion classes of natural language and a scenario for its use. Sentiment analysis is a binary classification task whereas the affect sensing is a multi-class problem. A new approach has been demonstrated the use of large-scale real-world knowledge about the inherent affective nature of everyday situations (such as "getting into a car accident") to classify sentences into "basic" emotion categories. Open Mind[5] Commonsense knowledge has been used as a real world corpus of 400,000 facts about the everyday world. Four linguistic models (Statistical-Syntactic) are combined for robustness as a society of commonsense-based affect recognition. The results suggest that the approach is robust enough to enable plausible affective text user interfaces for future use. This work is also very important in another aspect as it shows the possibility that contextual polarity could be inferred by the syntactic formulations and a formidable accuracy could be reached by this method. The syntactic-statistical techniques for the polarity classification problem have been attempted in several works with good accuracy (Seeker et al., 2009; Moilanen et al., 2010).

Sentiments of people are important because people's sentiment has great influence on our society. But knowing only the positive or negative aspect of sentiments is not enough because the end users of the proposed Sentiment Analysis systems might look for the comparative or evaluative study for making their own decisions. For example, we always look for a feature wise comparative study before buying any product (Is the product X better than product Y?) or before casting our vote for any candidate (Is Mr. X better than Mr. Y?). To meet such real life necessities, (Liu et. al., 2005) developed a system called

[5] http://www.openmind.org/

Opinion Observer that can analyze and compare opinions available on the Web. The system is such that with a single glance of its visualization, the user is able to clearly see the strengths and weaknesses of the various features of each product in the minds of consumers. A technique based on language pattern mining has been proposed to extract Pros (positive) and Cons (negative) product features in a particular type of reviews. Experimental results show that the technique is highly effective and it outperforms existing methods significantly.

Following the same line of hypothesis as (Liu et. al., 2005), (Pang and Lee, 2005) have proposed a sentiment rating technique by viewing the number of stars provided by each customer to each product from customer feedback. The sentiment rating task has been described as a multi-class problem. The standard machine learning technique, Support Vector Machine (SVM) has been used in this setup.

The above mentioned sentiment categorization tasks make an implicit assumption that a single score can express the polarity of an opinion text. However, multiple opinions on related matters are often intertwined throughout a text. For example, a restaurant review may express judgment on food quality as well as the service and ambience of the restaurant. Rather than accumulating these aspects into a single score, people may get interested to know the aspectual sentiment separately. Therefore to provide such facility, (Snyder and Barzilay, 2007) have proposed their Multiple Aspect Ranking technique using the Good Grief Algorithm. The Good Grief algorithm guides the prediction of individual rankers by analyzing meta-relations between opinions, such as agreement and contrast. Probably this is the first attempt when researchers started using data mining based semantic association models for polarity classification task. This kind of modeling has been attempted by other researchers later (Speriosu et. al., 2011).

The Text Retrieval Conference (TREC) Polarity Classification of Blog track[6] 2008 brought together researchers to share their knowledge and compare the efficiency of their proposed techniques. Most of the submitted runs for the task used a two-stage approach (prior polarity identification followed by NLP or other techniques). Only 12 runs out of the submitted 191 runs did not adopt this strategy. The three opinion-finding approaches out of these 12 runs that were consistently effective across the entire provided baseline have been focused in the present work.

The approach by University of Illinois at Chicago (Jia et al., 2008) achieved the best average improvement over the standard topic-relevance baselines (an average of 11.76% improvement) for the opinion-finding. Sentence level and document level polarity classification models have been developed and finally the polarity scores have been accumulated to generate the final result. The sentence level classifier is a simple SVM based classifier that classifies a query relevant opinion sentence as either positive or negative. Two approaches were proposed at document level, a Heuristic Rule Based Model and the Decision Tree Model. The Heuristic Rule Based polarity classification system was developed based on the following intuition: a document is positive (negative) if it only contains positive (negative) relevant opinions. If the document contains both kinds of opinions, it needs further analysis. If the

[6] http://trec.nist.gov/pubs/trec17/papers/BLOG.OVERVIEW08.pdf

positive (negative) relevant opinions are significantly stronger than the negative (positive) relevant opinions, the opinion polarity of this document should be positive (negative). The Decision Tree Model is a machine learning method that improves the document-level opinion polarity classification accuracy. A vector of polarized words/phrases is formed for each document whose polarity is determined initially by the sentence level classifier. This research effort is very significant because it shows the clear distinction between the sentence level and the document level polarity classification. Later, many other researchers (Somsundaram and Wiebe, 2009) have considered the polarity classification problems distinctly at sentence level and document level.

The approach by (Lee et. al., 2008) has used a domain specific lexicon based approach. In addition to SentiWordNet, the authors have used Amazon's product review corpus and product specification corpus to create the opinionated lexical resource. This clearly shows that domain knowledge is required for polarity classification along with generic prior polarity lexicons like SentiWordNet. The accuracy of a polarity classifier mainly depends on the handling of unknown words or new words. The same conclusion has been drawn by several other researchers later (Aue and Gamon, 2009; Takamura et. al., 2005) as well.

(He et. al., 2008) have used their domain specific divergence model for polarity classification task. The work is based on the hypothesis that the semantic orientation of prior polarity lexicon from a pre-processed dictionary may vary in the current domain. The authors have enhanced a dictionary-based approach by automatically building an internal opinion dictionary from the provided corpus collection itself. This approach measures the opinionated discrimination property of each term in the dictionary using information theoretic divergence measure based on the relevance assessments at context level.

3.2.3 Mimicking the Human Psychology to Solve the Sentiment Analysis

The Sentiment Analysis research has become quite matured after a few decades of research. As a result, a few systems like Twitter Sentiment Analysis Tool (http://twittersentiment.appspot.com/), TweetFeel (http://www.tweetfeel.com/) are available in the World Wide Web since last few years. More research efforts are necessary to meet the satisfaction level of the end users (Liu, 2010). The main issue is that there are many conceptual rules that govern sentiment and there are even more clues (possibly unlimited) that can convey these concepts from realization to verbalization of a human being. Human psychology may provide the unrevealed clues and govern the sentiment realization. Human psychology relates to social, cultural, behavioral and environmental aspects of civilization. The important issues that need attention include how various psychological phenomena can be explained in computational terms and the identification of the various Artificial Intelligence (AI) concepts and computer modeling methodologies that are most useful from the psychologist's point of view. An important research endeavor could be noticed supporting this notion in the form of a workshop "*Sentiment Analysis where*

AI meets Psychology (SAAIP 2011)[7]" held as part of the International Joint Conference on NLP (IJCNLP 2011).

(Cambria et al., 2011) did a wonderful contribution in this direction. They introduced a new paradigm, called Sentic Computing[8], in which an emotion representation and a Common Sense[9] (Cambria et al., 2009) based approach have been used to infer affective states from short texts over the web. The innovation of Sentic Computing is a result of in-depth scientific cultivation by several other researchers over two decades. Some of those important research attempts which made the avenue to the present Sentic computing are reported below.

The term '*sentic*' is derived from the Latin word '*sentire*', the root of words like sentiment and sensation. It was first adopted in 1977 by Manfred Clynes (Clynes, 1977), who discovered that when people have emotional experience, their nervous system always responds in a characteristic way which is measurable. Sentic Computing is part of the efforts in the fields of computer science, psychology, linguistics, sociology and cognitive science, to develop a kind of computing that relates to or arises from or influences emotions (Picard, 1997). The approach adopted by (Liu et. al., 2003) exploits a Common Sense knowledge base to extract affective information from emails using the standard notion of basic emotions provided by Ekman[10]. Nowadays, researchers use a much richer semantic network, ConceptNet[11] (Havasi et. al., 2007), with almost 10,000 concepts and a set of 72,000+ features extracted from the Open Mind corpus[12] along with the power of cumulative analogy provided by AnalogySpace, a process which reveals large-scale patterns in the data, smoothes over noise and predicts new knowledge.

The aim in Sentic Computing is to develop emotion-sensitive systems that can measure how much:

1. The user is happy with the service provided?
2. The user is interested in the information supplied?
3. The user is comfortable with the interface?
4. The user is keen on using the application?

Thus, in Sentic Computing the user's affective states are organized around four independent dimensions: *Pleasantness, Attention, Sensitivity and Aptitude*. This model is a variant of Plutchik's wheel of emotions (Plutchik, 2001) and constitutes an attempt to emulate Marvin Minsky's conception of

[7] http://saaip.org/
[8] http://cs.stir.ac.uk/~eca/sentics
[9] http://cs.stir.ac.uk/~eca/commansense
[10] http://en.wikipedia.org/wiki/Paul_Ekman#Emotion_classification
[11] http://conceptnet5.media.mit.edu/
[12] http://www.openmind.org/

emotions (***The Emotion Machine***[13]: Minsky, 2006). Minsky sees the mind as a collection of thousands of different resources and believes that our emotional states result from turning some set of these resources on and turning another set of them off. Each such selection identifies how we think by changing our brain's activities: the state of anger, for example, appears to select a set of resources that help us to react with more speed and strength while also suppressing some other resources that usually make us act prudently.

Another important contribution in this direction is by (Grassi, 2009). The Human Emotion Ontology (HEO) is conceived as a high level ontology for human emotions that supplies the most significant concepts and properties which constitute the centerpiece for the description of every human emotion. The main purpose of HEO and Sentic Computing is to create a descriptive lexical network framework that can hold the sentimental/emotional and common sense knowledge with enough flexibility, by allowing the use of a wide and extensible set of descriptors to represent all the main features of an emotion and with interoperability, by allowing to map concepts and properties belonging to different emotion representation models. For example, HEO is visually represented in the Figure3.1.

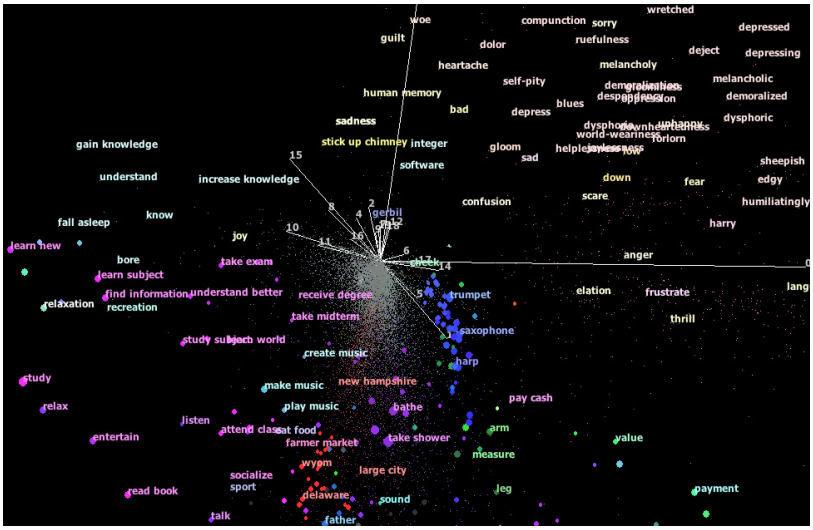


Figure 3.1: Affective Space in Human Emotion Ontology (Grassi, 2009)

The Sentic Computing team trusted on Hourglass of Emotions visualization which is a novel affective categorization model in which sentiments are organized around four independent dimensions, whose different levels of activation constitute the total emotional state of the mind. The Hourglass of Emotions, in fact, is based on the idea that the mind is made of different independent resources and that emotional states result from turning some set of these resources on and turning another set of them off (Minsky, 2006). An example of Hourglass of Emotions visualization is shown in the Figure 3.2.

[13] http://web.media.mit.edu/~minsky/Introduction.html

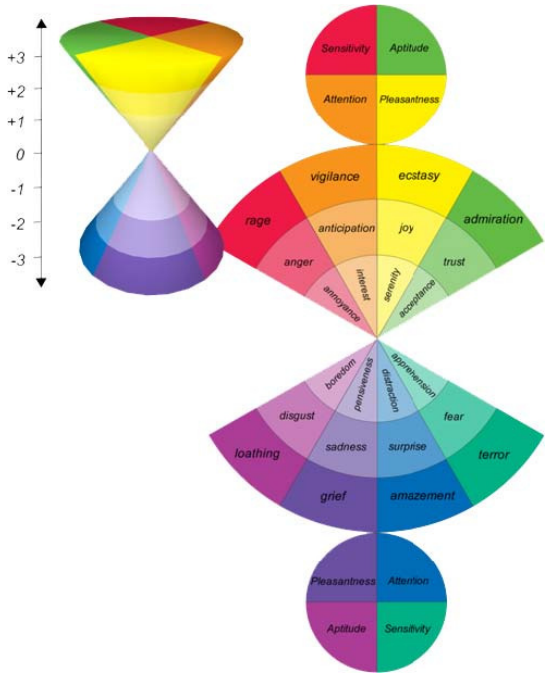


Figure 3.2: The Hourglass of Emotions (Cambria et al., 2011)

3.3 Resource Acquisition

In the present work, the polarity classification experiments have been carried out for Bengali language. The Sentiment Analysis task for a new language demands linguistic resources like gold standard annotated data and other NLP tools. The basic polarity classification task in the present work started with syntactic dependency (Liu et. al., 2003). Therefore, a dependency parser is necessary for the experiments. Bengali is a resource scarce language and no Bengali Dependency parser was available when the work started. Thus, the development of Bengali Dependency parser was identified an important task for the present work. The details of the development of the annotated corpora and Dependency parser are described in the sections 3.3.1 and 3.3.2 respectively.

3.3.1 Corpus

All the experiments in the present work have been carried out on Bengali news corpus. News text can be divided into two main types: (1) news reports that aim to objectively present factual information and (2) opinionated articles that clearly present authors' and readers' views, evaluation or judgment about some specific events or persons. Type (1) is supposed to be the common practice in newspapers, and Type (2) appears in sections such as 'Editorial', 'Forum' and 'Letters to the editor'. 'Reader's opinion' section or 'Letters to the Editor' Section from the web archive of a popular Bengali newspaper[14] have

[14] http://www.anandabazar.com/

been identified as the relevant corpus in Bengali. A brief statistics about the corpus have been reported in the Table 3.2. The corpus is then manually annotated. The annotation scheme used in the corpus annotation is reported in Figure 3.3. The positive algebraic sign in the feature structure ("<fs af=,**+**,") depicts the phrase polarity as positive and the negative algebraic sign in the feature structure ("<fs af=,**_**,") depicts the phrase polarity as negative .

Corpus Statistics	
Total number of documents in the corpus	20
Total number of sentences in the corpus	447
Avgerage number of sentences in a document	22
Total number of wordforms in the corpus	5761
Avgerage number of wordforms in a document	288
Total number of distinct wordforms in the corpus	3435

Table 3.2: Statistics on Bengali Polarity Annotated News Corpus

```
2       ((          CCP
2.1     যেমন        CC
3       ((          NP      <fs af=',+,,,,,,' name='?'>
3.1     মঙ্গলজনক            NN
```

Figure 3.3: Bengali Corpus Polarity Annotation Scheme

3.3.2 Dependency Parser

To build the Dependency Parser we participated in the ICON 2009[15] and 2010[16] NLP TOOLS CONTEST: IL Dependency Parsing tasks. The input files provided as part of the contest are in the Shakti Standard Format (SSF)[17] that includes the POS tags, Chunk labels and morphology information. A probabilistic sequence model has been followed which allows integrating uncertainty over multiple, interdependent classifications and collectively determines the most likely global assignment. Standard machine learning models like, Conditional Random Field (CRF) and Support Vector Machine (SVM) have been used. The chunk information in the input files are converted to B-I-E format so that the begin (B) / inside (I) / End (E) information for a chunk are associated as a feature with the appropriate words. The chunk tags in the B-I-E format of the chunk with which a particular chunk is related through a dependency relation are

[15] http://ltrc.iiit.ac.in/nlptools2009/
[16] http://ltrc.iiit.ac.in/nlptools2010/
[17] http://www.docstoc.com/docs/7232788/SSF-Shakti-Standard-Format-Guide

identified from the training file and noted as the input features in the machine learning (ML) based system. The corresponding relation name is also another input feature associated with the particular chunk. Each sentence is represented as a feature vector for the ML based machine learning task. After a series of experiments the following feature set is found to be performing well as a dependency clue. The input features associated with each word in the training set are the root word, pos tag, chunk tag and vibhakti or the inflection. The details of the development process could be found in (Ghosh et. al., 2009) ; (Ghosh et. al., 2010) appendix section.

3.4 The Syntactic Polarity Classifier

The two step methodology, i.e., use of prior polarity lexicon followed by any NLP technique is the standard method for the polarity classification task. The Bengali SentiWordNet [18] developed as part of the present work and discussed in the Chapter One has been used as the prior polarity lexicon. For the NLP technique, the Syntactic-Statistical classification NLP technique has been used (Das and Bandyopadhyay, 2010(a));(Das and Bandyopadhyay, 2010(h)). The syntactic clue directly helps to understand the relation between the localized semantic orientation, i.e., word level semantic orientation and the contextual semantic orientation, i.e., word/phrase/sentence level semantic orientation. In the following example sentence, the localized semantic orientation at word level, ভালো (good) could be obtained directly from the prior polarity lexicon as positive.

He is not a good^{+} boy.

সে ভালো$^{+}$ ছেলে নয়।

The negation word 'not' changes the contextual semantics in the opposite direction, i.e., negative. To understand this contextual feature, the syntactic relationship helps as the word "not (নয়)" has a modifier relationship with the word "good (ভালো)" (modified). Therefore, it is very easy to infer the resultant contextual semantic orientation of the sentence as negative.

Moreover the syntax sometime helps to predict the semantic orintation of any new word. Let us take a look at the following example sentence.

This is ugly^{-} and smelly.

এটি বিশ্রী$^{-}$ এবং কটুগন্ধযুক্ত।

Let us consider that the prior polarity lexicon only covers *"ugly (বিশ্রী)"* and not the *"smelly (কটুগন্ধযুক্ত)".* As the semantic orientation of the word *"ugly (বিশ্রী)"* is negative it is more or less obvious that the semantic orientation of the new word *"smelly (কটুগন্ধযুক্ত)"* will be the same because it has been

[18] http://www.amitavadas.com/sentiwordnet.php

observed that generally words with same orientation are syntactically joined with "and" and words with orthogonal semantic orientation are syntactically joined with "but/rather/either…etc" as seen in the following example sentence.

Good^{+} but costly^{-}.

ভালো$^{+}$ কিন্তু দামী$^{-}$।

Several other researchers (Liu et. al., 2003; Seeker et. al., 2009; Moilanen et. al., 2010) have also identified the same linguistic phenomena. But there are exceptions like, "*The Good Bad and Ugly"*. In the famous movie title, "*Bad and Ugly*" is syntactically joined by the conjunct "*and*" with the word "*Good*". The three adjectives in the title metaphorically refer to three entities or three persons who are the characters in the movie. Such exceptions are rare in the language.

It has also been observed that localized syntax helps to understand the discourse level sentimental semantics to some extent (Somsundaram, 2009). For example, the following sentences are from two different paragraphs from the same document.

The reason behind the electoral disaster is the wrong **policy of the previous Government**.

পূর্বতন সরকারের ভুল নীতি ভোটে ভরাডুবির অন্যতম কারন।

We will not follow the **strategy of the previous government**, said Mamata Banerjee.

মমতা বানার্জী বলেন আমরা পূর্বতন সরকারের নীতি অনুসরন করবো না।

In the first sentence the word "*wrong (ভুল)*" is modifying the phrase "*strategy of the previous Government (পূর্বতন সরকারের নীতি)*" and it is negative. Therefore in the same scope of the document it is very likely that a single author will not sentimentally differ too much regarding the same topic and thus the final semantic orientation of the second sentence is likely to be positive as it includes a negation. But it is very hard to assimilate this kind of knowledge into the Syntactic-Statistical polarity classifier. An in-depth semantic tagging at the discourse level is required for this kind of work.

3.4.1 Features Extraction

The standard machine learning method Support Vector Machine (SVM)[19] has been used for the present syntactic statistical polarity classifier. The SVM has a few advantages over the other existing machine learning techniques that depend on the data being analyzed. The typical scenario for the SVM is when the data are not regularly distributed or have an unknown distribution. Sentiment analysis data is a perfect example of this type. No one can predict in which order the positive or negative words will occur in a text, i.e., there is no regular distribution. It completely depends on the psychological forces of

[19] http://www.lsi.upc.edu/~nlp/SVMTool/

the situation that were in effect when the document was written by a particular writer. Moreover sentiment is not a linguistic phenomena and it is nearly impossible to identify the concrete set of psychological or cognitive features from the written text. A detailed psycho-linguistic study is necessary which demands more and more reliable linguistic tools but unfortunately such tools are unavailable for Bengali language. SVM works well with less numbers of distinct informative features which is essential for working with a new language. SVM provides a good out-of-sample generalization. It means that, by choosing an appropriate generalization grade, SVMs can be robust, even when the training sample has some bias or limitations (Auria and Moro, 2008). To support the argumentation in favor of SVM, experiments were conducted using the CRF machine learning technique with the same data and setup. The comparative results are reported in the Section 3.4.2 and it proves the effectiveness of using SVM in the current setup.

SVM treats opinion polarity identification as a sequence tagging and pattern-matching task, acquiring symbolic patterns that rely on both the syntax and lexical semantics of a phrase and sentence. Several word level features are extracted using different tools from the input sentences. The feature identification starts with Part Of Speech (POS) categories and the exploration is continued with other features like chunk, functional word, SentiWordNet (Bengali), stemming cluster, Negative word list and Dependency tree features. The feature extraction for any Machine Learning task is crucial since proper identification of the entire features directly affects the performance of the system. Functional word, SentiWordNet (Bengali) and Negative word list features are fully dictionary based. On the other hand, POS, chunk, stemming cluster and dependency tree features are extractive.

3.4.1.1 Part Of Speech (POS)

It has been shown (Hatzivassiloglou et. al., 2000; Chesley et. al., 2006) that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion-topic identification systems, like (Nasukawa et. al., 2003) are based on adjective or adverb words. The Bengali Shallow Parser[20] developed under the "Indian Languages to Indian Languages machine Translation (IL-ILMT)" project funded by Department of Information Technology; Government of India has been used in the present work.

3.4.1.2 Chunk

In the Syntactic-Statistical polarity classifier local dependencies like chunk boundaries and chunk member information are very important features. It is not unusual for two annotators to identify the same expression as a polar element in the text, but they could differ in how they mark the boundaries, such as the difference between '*such a disadvantageous situation*' and '*such…disadvantageous*' (Wilson and Wiebe, 2003). Similar fuzziness appeared in the marking of polar elements in the present task, such as 'কেন্দ্রীয় দলের দুর্নীতিতে' (corruption of central team) and 'দূর্নীতিতে' (corruption). Hence the hypothesis is

[20] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

to stick to the automatically assigned chunk labels only to avoid any further ambiguity. Chunk level information is effectively used as a feature in the supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label.

3.4.1.3 Functional word

Function words in a language are high frequency words and these words generally do not carry any opinionated information. But function words help many times to understand the syntactic pattern of a sentimental text. A list of 253 functional words is collected from the Bengali corpus. First a unique high frequency word list is generated where the assumed threshold frequency is considered as 20. Then the list is manually corrected. The function word feature is very important to disambiguate the contextual polarity for unknown words.

3.4.1.4 Prior Polarity Lexicon

The classical two step methodology for the polarity classification problem, i.e., prior polarity lexicon followed by NLP techniques for further contextual polarity disambiguation, has been followed in the present work. The developed Bengali SentiWordNet is used as the prior polarity lexicon in the present work. Words that are present in the SentiWordNet carry sentiment information. The prior polarity lexicon features are individual sentiment words or word n-grams (multiword entities) with polarity values either positive or negative. Positive and negative polarity measures are treated as a binary feature in the supervised classifier. Words which are collected directly from the SentiWordNet are tagged with positivity or negativity scores.

3.4.1.5 Stemming cluster

Several words in a sentence that carry opinion information may be present in inflected forms. Stemming is necessary for such inflected words before they can be searched in the appropriate lists. Due to non availability of good stemmers in Indian languages, especially in Bengali, a stemmer based on stemming cluster technique has been evolved. This stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as the cluster center. Details can be found in (Das and Bandyopadhyay, 2010(l)).

3.4.1.6 Negative words

Negative words like no (না), not (নয়) etc. does not carry any sentiment information but these words relationally affect the resultant polarity of any polar phrase. A manually edited list of negative words has been used as a binary feature in the SVM classifier.

3.4.1.7 Dependency Tree feature

Dependency relations are the most crucial feature in the Syntactic-Statistical polarity classifier. The feature extractor module searches the dependency tree using the breadth-first search technique to identify syntactically related nodes and their mutual relations. The purpose of the feature is to encode dependency structure between related polar phrases. The details of the dependency parser as used in the present task have been described in the Section 3.3.2.

3.4.2 Performance of the Syntactic Polarity Classifier

The evaluation result of the SVM-based polarity classification task for Bengali is presented in Table 3.3. The evaluation results of the system for the positive and negative polarity classes are mentioned separately in Table 3.4.

Language	Precision	Recall
Bengali	70.04%	63.02%

Table 3.3: The Overall Performance of Polarity Classification for Bengali

Polarity	Precision	Recall
Positive	56.59%	52.89%
Negative	75.57%	65.87%

Table 3.4: Polarity Wise Performance of Polarity Classification for Bengali

To understand the effects of various features on the performance of the system the feature ablation method has been studied. The dictionary based approach using only the SentiWordNet has accuracy (precision) of 47.60% and this may be considered as the baseline system. It may be observed from the Table 3.5 that incremental use of other features like negative word, functional word, parts of speech, chunk and tools like stemming cluster has improved the precision of the system to 66.8%. Thus an increase of 19.2% in precision over the baseline system has been obtained. Further use of syntactic feature in terms of dependency relations has improved the system precision to 70.04%. Thus an increase of 3.6% in precision has been obtained due to the use of syntactic feature. The feature ablation method proves the effectiveness of the two step polarity classification technique. The prior polarity lexicon, i.e., completely dictionary based approach produces 47.60% precision and further improvement of the system could be achieved using various NLP techniques. The importance of each feature has been identified with the feature ablation method and is shown in Table 3.5.

To support the arguments for choosing SVM machine learning method, the same classification problem was attempted using CRF machine learning technique with the same data and setup. The resulting accuracy of the CRF based model with precision 61.23% and recall 55.0% is much less than the SVM based model. The same feature ablation method as reported in the Table 3.5 was applied on the CRF based model. It has been noticed that the accuracy level is more or less same till the dictionary features

and lexical features (SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech) are used. But it is hard to increase the performance level of the CRF based model when the syntactic features like chunk and dependency relations are used. SVM machine learning technique works excellent to normalize this dynamic situation.

Features	Performance
SentiWordNet	47.60%
SentiWordNet + Negative Word	50.40%
SentiWordNet + Negative Word + Stemming Cluster	56.02%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word	58.23%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech	61.9%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech +Chunk	66.8%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech + Chunk +Dependency tree feature	70.04%

Table 3.5: Performance of the Syntactic Polarity Classifier by Feature Ablation

It may be suggested in the present situation that a multi-engine based voting technique could work well because it has been noticed that such methods work well for this type of heterogeneous tagging task like NER (Ekbal and Bandyopadhyay, 2010) and POS tagging (Shulamit et al., 2010).

3.5 What Knowledge to Keep at What Level?

The use of the prior polarity lexicon has proved itself as a strong baseline in the polarity classification task. The feature ablation method as reported in the earlier section has strongly supported the requirement for further NLP techniques.

Dealing with unknown/new words is a common problem in NLP tasks. It becomes more difficult for sentiment analysis because it is very hard to find out any contextual clue to predict the sentimental orientation on any unknown/new word. There is another problem of word sense disambiguation (Cem et. al., 2011), which is indeed a significant subtask when applying a resource like SentiWordNet. A prior polarity lexicon is attached with two probabilistic values, i.e., positivity and negativity scores but there is no clue in the SentiWordNet regarding ***which value to pick in what context?*** The general trend is to pick the highest one but that may vary with context. The following example may illustrate the problem better: the word "***High***" (Positivity: 0.25, Negativity: 0.125 for "***High***" in the SentiWordNet) is attached

with a positive (positivity value is higher than the negativity value) polarity in a text but the polarity of that word may vary in any particular use. The word "high" has a positive polarity in the first sentence while the same word has a negative polarity in the second sentence.

Sensex reaches high^{+}.

Price goes high^{-}.

Actually further NLP techniques are required to disambiguate these types of words. The statistics from the SentiWordNet (English) is presented in Table 3.6 to understand the big picture that shows how many words are ambiguous and need a special care. There are 6619 lexicon entries in the SentiWordNet where both the positivity and the negativity values are greater than zero whereas the total number of entries in the SentiWordNet (English) is 115424. Therefore, these entries are ambiguous because there is no clue in the SentiWordNet ***which value to pick in what context?*** Similarly there are a total of 17927 lexical entries in the SentiWordNet, whose positivity and negativity value difference is less than 0.2. These are also the ambiguous words.

Type	Number
Total Token	**115424**
Positivity>0 && Negativity>0	**6619**
Positivity>0 || Negativity>0	**28430**
Positivity>0 && Negativity=0	**10484**
Positivity=0 & Negativity>0	**11327**
$\lvert Positivity - Negativity \rvert$>=0.2	**17927**

Table 3.6: A Closer Look on the Ambiguous Entries of SentiWordNet

The research attempts in the present work mainly concerns the ambiguous entries in the SentiWordNet. The basic hypothesis is that if we can add some contextual information along with the prior polarity scores in the sentiment lexicon, the updated rich lexicon network will serve better than the existing one and it may lessen the requirement of further NLP techniques to disambiguate the contextual polarity. A new paradigm called ***Sentimantics*** has been introduced which can be defined as the Distributed Semantic Lexical Model to hold the sentiment knowledge with contextual common sense. The new paradigm has some ideological similarity with the research attempt by (Erik et. al., 2011) as mentioned in the Section 3.2.3 but the intensions are totally different. The motivation of (Erik et. al., 2011) was to develop a four-dimensional vector representation for affect computing but the intension in the new paradigm called "Sentimantics" is to develop a rich lexical network of sentiment knowledge with contextual common sense that can be easily extractable. Moreover, the vector similarities within such semantic spaces have been shown to substantially correlate with human similarity judgments (McDonald, 2000) and word association norms (Denhire and Lemaire, 2004).

3.6 The Sentimantics and It's Motivation

To overcome the problems of the present proximity based static sentiment lexicon based techniques, a new way has been introduced to represent sentiment knowledge in a Vector Space Model (VSM) model. The proposed new models can store dynamic prior polarity with its different contextual information whereas the present prior polarity lexicons are static and has no contextual information. The representation of the sentiment knowledge in the Conceptual Spaces of Semantics is defined as ***Sentimantics*** (Das and Bandyopadhyay, 2012(c)).

To give sentimental cognition to the emotionally challenged machines we have to mimic the fundamentals of human cognitions properties. A relatively rich lexicon with context is required for this scenario and the lexicon should be organized as the mental lexicon or should have the contextual common sense properties. Similar lexical resources like ConceptNet[21], MindNet[22] or other works of Conceptual Spaces of Semantics are found in the literature. Therefore, the motivation is not new but the idea has not been used for the sentiment analysis task before. One of the fundamental problems of lexical semantics is the fact that the "*perceived meaning*" of a word can vary so greatly from one context to another (C Ruhl, 1989). Vector-based models (VSM) of word meaning (Lund and Burgess, 1996; Landauer and Dumais, 1997) have become increasingly popular in natural language processing (NLP) and cognitive science. The appeal of these models lies in their ability to represent meaning simply by using distributional information under the assumption that words occurring within similar contexts are semantically similar (Harris, 1954). The idea of the VSM is to represent each document in a collection as a point in a space (a vector in a vector space). Points that are close together in this space are semantically similar and points that are far apart are semantically distant. The success of the VSM for information retrieval has inspired researchers to extend the VSM to other semantic tasks in natural language processing with impressive results.

VSMs have several attractive properties. VSMs automatically extract knowledge from a given corpus using unsupervised techniques, thus they require much less labor than other approaches to semantics, such as hand-coded knowledge bases and ontologies. Vectors are common in AI and cognitive science; they were common before the VSM was introduced by (Salton et. al., 1975). The novelty of the VSM was to use frequencies in the text corpus as a clue for discovering semantic information. In cognitive science, Latent Semantic Analysis (LSA) (Deerwester et. al., 1990; Landauer and Dumais, 1997), Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996), and related research (Landauer, McNamara, Dennis, and Kintsch, 2007) are entirely within the scope of VSMs, since the research uses vector space models in which the values of the elements are derived from event frequencies, such as the number of times a given word appears in a given context. Cognitive scientists have argued that there are empirical and theoretical reasons for believing that VSMs, such as LSA and HAL, are plausible models of some aspects of human cognition (Landauer et. al., 2007). In AI, computational linguistics, and information

[21] http://csc.media.mit.edu/conceptnet
[22] http://research.microsoft.com/en-us/projects/mindnet/

retrieval, such plausibility is not essential, but it may be seen as a sign that VSMs are a promising area for further research.

3.7 Technical Solutions for Sentimantics

Two different type models for Sentimantics composition have been examined that are empirically grounded and can represent the contextual similarity relations among various lexical sentiment and non-sentiment concepts. The work on proposing models for Sentimantics composition started with Semantic Network Overlap Technique with the existing resources like ConceptNet and SentiWordNet for English and SemanticNet (Das and Bandyopadhyay, 2010(p));(Das and Bandyopadhyay, 2010(n)) and SentiWordNet (Bengali)[23] for Bengali. We call this as a Semantic Network Overlap Technique. The common sense lexicons like ConceptNet and SemanticNet have been developed for general purpose. The formalization of Sentimantics from these resources faces problems due to lack of dimensionality. Section 3.7.1.2. presents more rational explanations with empirical results. Therefore, a VSM has been developed to hold the Sentimantic from scratch by a corpus driven semi-supervised method. This model relatively performs better than the previous one, i.e. the Semantic Network Overlap Technique. Extraction of knowledge from this kind of VSM is generally very expensive because it is a very high dimensional network. Another important limitation of this type of model is that it demands very well defined processed input to extract knowledge like: *Input: (high) Context (sensex, share market, point)*, which demands NLP pre-processing techniques for the input text to extract knowledge from this VSM. Philosophically, the motivation of Sentimantics is to provide a rich lexicon network that will serve well than the existing one and it may lessen the requirement of further NLP techniques to disambiguate the contextual polarity. Therefore, the Syntactic Co-Occurrence Based VSM with relatively fewer dimensions has been proposed. The final model is the best performing lexicon network model for the Sentimantics problem. The details of the proposed models are described in the following sub-sections.

3.7.1 Starting with Existing Resources: Semantic Network Overlap

Several common sense networks are publicly available in the Web, for example, the ConceptNet for English and the SemanticNet (Das and Bandyopadhyay, 2010(p));(Das and Bandyopadhyay, 2010(n)) for Bengali. The experiments started with the network overlap technique which finds overlaps of nodes between two lexical networks, i.e., ConceptNet-SentiWordNet (English) and SemanticNet-SentiWordNet (Bengali). The working principle of the network overlap technique is very simple. The algorithm starts with any SentiWordNet node and then finds its close neighbors from the commonsense networks like ConceptNet/SemanticNet. For example, the node chosen from SentiWordNet, "long/লম্বা", has its close neighbors along with the context information extracted from the commonsense networks: "*road* (40%)/*waiting* (62%)/*car* (35%) /*building* (54%) /*queue* (70%) …" and "লাইন (66%)/দিন (46%)/প্রতীক্ষা (75%)...". The association scores (as in the previous example) are extracted to understand the semantic

[23] http://www.amitavadas.com/sentiwordnet.php

similarity association. The next prime challenge is to assign contextual polarity to each association. To associate lexical contextual polarity, a corpus based method has been followed. The Multi Perspective Question Answering (MPQA) corpus has been chosen for English and the experiments in Bengali have been carried out with the news corpus developed as part of the present work (as described in the section 3.3.1). The corpus is pre-processed with Dependency relations and stemming. The dependency relations are necessary to understand the relations between the evaluative expressions and other modifier-modified chunks in any subjective sentence. The Stanford Dependency Parser for English and the Bengali Dependency Parser (as discussed in the section 3.3.2) has been used in the present task. Stemming is necessary to identify the root form of any word so that it can be compared with the dictionary entries. The Porter stemmer for English and the Stemming Cluster based technique for Bengali, described in appendix section (Das and Bandyopadhyay, 2010(l)) have been used in the present task.

By the corpus driven method each sentiment word in the developed lexical network is assigned a contextual prior polarity by the corpus driven method. The lexical network for the word "*long*" is shown in the Figure 3.4.

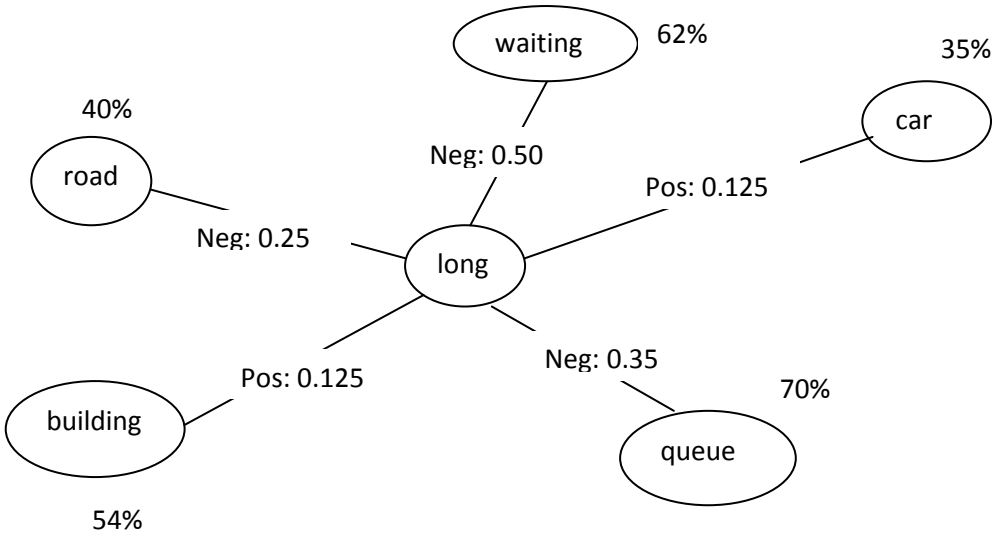


Figure 3.4: The Developed Lexical Network by Network Overlap Technique

3.7.1.1 Polarity Identification from the Semantic Network Overlap

Once the desired lexical semantic network to hold the Sentimantics has been developed, the developed knowledge is used further for the polarity classification task. The motivation of Sentimantics is to provide a rich lexicon network that will serve well than the existing ones and it may lessen the requirement for further NLP techniques to disambiguate the contextual polarity. The methodology of contextual polarity extraction from this network is very simple. For this task only a dependency parser and stemmer are required. Let us consider the following example sentence.

We have been waiting in a long queue.

To extract the contextual polarity from the given sentence it is necessary to know that *waiting-long-queue* is interconnected with dependency relations (*modifier-modified*) and stemming is a necessary pre-processing step for dictionary matching. The contextual polarity can be extracted from the developed network with the input (*long*) along with its context (*waiting, queue*). The accumulated contextual polarity will be Neg: (0.50+0.35) =0.85. For comparison, the score extracted from the SentiWordNet (English) is Pos: 0.25 as the word "long" has the entry (long: Pos: 0.25 Neg: 0.125) in the SentiWordNet (English)) and the positive score is greater than the negative score.

3.7.1.2 Performance of the Semantic Network Overlap and the Limitation

Enriching the existing fixed point prior polarity technique can help to lessen requirements of further NLP techniques to disambiguate the contextual polarity. The Sentimantics lexicon resource with contextual prior polarity score has been developed for polarity classification. The Sentimantics developed using the Network Overlap technique outperforms previous lexical resources for the polarity classification task. The reported accuracy of Sentimantics based polarity classification task is 62.3% for English and 59.70% for Bengali using the MPQA and Bengali corpus developed as part of the present work. The scores are comparatively higher than the SentiWordNet based baseline system which is 47.60% as reported in Table 3.5.

Type	Number	Solved By Semantic Network Overlap Technique
Positivity>0 && Negativity>0	6619	**2304**
$\lvert Positivity - Negativity \rvert$ >=0.2	17927	**5230**

Table 3.7: Result of the Semantic Network Overlap Technique

During error analysis and identification of the missed cases of the present system, the coverage was found as the main issue. Both the ConceptNet and the SemanticNet have been developed from the news domain but for different tasks. The comparative coverage of the SentiWordNet (English) and MPQA corpus is 74%, i.e., 74% of the complete set of sentiment words from MPQA corpus are covered by the SentiWordNet (English). This is a very good and acceptable coverage. For Bengali the comparative coverage is 72%, which is also very good. But the comparative coverage of SentiWordNet (English)-ConceptNet and SentiWordNet (Bengali)-SemanticNet are as low as 54% and 49.6% respectively. It means that only 54% of sentiment words from SentiWordNet (English) have been covered by the ConceptNet and only 49.6% of sentiment words from SentiWordNet (Bengali) have been covered by the SemanticNet. Table 3.7 reports the performance of the polarity classification task in English based on the proposed Semantic Network Overlap based technique. The results are not satisfactory because only 34% cases of "*Positivity>0 && Negativity>0*" has been resolved and only 30% cases of

"$|Positivity - Negativity|$>=0.2" has been resolved by this technique. The result presented in the Table 3.7 is for English.

After error analysis, it was decided to develop the Vector Space Model (VSM) from scratch to solve the Sentimantics issue in order to reach a satisfactory level of coverage.

3.7.2 Starting from Scratch: Syntactic Co-Occurrence Network Construction

A syntactic word co-occurrence network has been constructed only for the sentimental words from the MPQA (Multi Perspective Question Answering)[24] corpus. The syntactic network has been defined in a way similar to the Spin Model (Takamura et. al., 2005) or the Latent Semantic Analysis (Turney and Litman, 2003) to compute the association strength of words with seed words. The hypothesis is that all the words that occur in the similar syntactic territory tend to have similar semantic orientation. In the present work, only words with Noun, Verb, Adjective and Adverb POS categories have been considered to construct the network as these are open POS classes of words and tend to have maximum sentiment properties. Another vital reason is low dimensionality. Involvement of less number of features will generate VSM with fewer dimensions.

The network creation started with SentiWordNet 3.0 to mark the sentiment words in the MPQA corpus. As the MPQA corpus is marked at expression level, the SentiWordNet has been used to mark only the lexicon within the marked subjective expressions in the corpus. Stanford POS tagger[25] and Porter Stemmer[26] have been used to get the POS classes and the stems of a lexeme respectively.

A word window of ±4 words around the target words has been considered as the main feature. Clustering techniques have been used for the in depth analysis of word-co occurrence pattern and their relationship at discourse level. The clustering algorithms partition a set of lexicons into finite number of groups or clusters in terms of their syntactic co-occurrence relatedness.

The similarity between vectors is calculated by assigning numerical weights to these words and then using the cosine similarity measure as specified in the following equation.

$$s\left(\vec{q}_k, \vec{d}_j\right) = \vec{q}_k . \vec{d}_j = \sum_{i=1}^{N} w_{i,k} \times w_{i,j} \text{ ----- (3.1)}$$

This equation specifies what is known as the dot product between vectors. Of course, in situations where the lexicon collection is relatively static, it makes sense to normalize the vectors once and store them, rather than include the normalization in the similarity metric.

[24] http://www.cs.pitt.edu/mpqa/
[25] http://nlp.stanford.edu/software/tagger.shtml
[26] http://tartarus.org/martin/PorterStemmer/java.txt

$$s\left(\vec{q}_k, \vec{d}_j\right) = \frac{\sum_{i=1}^{N} w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,k}^2}} \quad \text{----- (3.2)}$$

Calculating the similarity measure and using a predefined threshold value, lexicons are classified using standard *k-means* clustering technique. The predefined threshold value is experimentally set as 0.5 as shown in Table 3.8.

A set of initial cluster centers is necessary in the beginning. Each document is assigned to the cluster whose center is closest to the document. After all documents have been assigned, the center of each cluster is recomputed as the centroid or mean $\vec{\mu}$ (where $\vec{\mu}$ is the clustering coefficient) of its members that is $\vec{\mu} = \left(1/\left|c_j\right|\right)\sum_{x \in c_j} \vec{x}$. The distance function is the **cosine vector** similarity function.

ID	Word	Cluster 1	Cluster 2	Cluster 3
1	Broker	0.63	0.12	0.04
1	NASDAQ	0.58	0.11	0.06
1	Sensex	0.58	0.12	0.03
1	*High*	0.55	0.14	0.08
2	India	0.11	0.59	0.02
2	Population	0.15	0.55	0.01
2	*High*	0.12	0.66	0.01
3	Market	0.13	0.05	0.58
3	Petroleum	0.05	0.01	0.86
3	UAE	0.12	0.04	0.65
3	*High*	0.03	0.01	0.93

Table 3.8: Syntactic Co-occurrence Lexical Network: Three Cluster Centroids

Table 3.8 gives an example of cluster centroids by the *K-means* clustering. Bold words in the column are cluster centers. Comparing two members of the cluster2, '**India**' and '**Population**', it is seen that '**India**' is strongly associated with cluster2 ($\vec{\mu}$=0.59) but it has some affinity with other clusters as well (e.g., p =0.11 with the cluster1). This is a good example of the utility of soft clustering. These non-zero values are still useful for calculating vertex weight during Semantic Relational Graph generation.

3.7.2.1 Polarity Calculation using the Syntactic Co-Occurrence Network

The relevance of the semantic lexicon nodes have been computed by summing up the edge scores of those edges that connect the node with other nodes in the same cluster. As cluster centers are also interconnected with weighted vertex so inter-cluster relations can also be calculated in terms of

weighted network distance between two nodes within two separate clusters. Let us consider the following example Semantic Affinity graph (Figure 3.5) for contextual prior polarity:

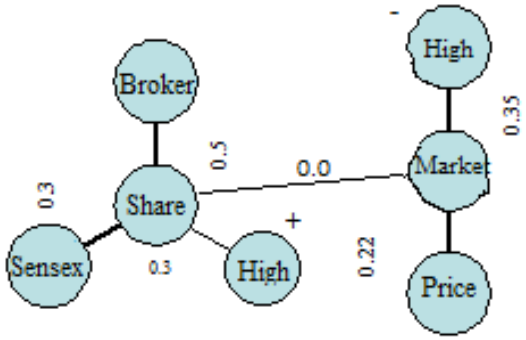


Figure 3.5: Semantic Affinity Graph for Contextual Prior polarity

The lexicon level semantic orientation from Figure 3.5 can be calculated as follows:

$$S_d(w_i, w_j) = \frac{\sum_{k=0}^{n} v_k}{k} * w_j^{\,p} \quad \text{-----} \quad (3.3) \text{ or}$$

$$= \sum_{c=0}^{m} \frac{\sum_{k=0}^{n} v_k}{k} * \prod_{c=0}^{m} l_c * w_j^{\,p} \text{-----} \quad (3.4)$$

where $S_d(w_i, w_j) =$ semantic orientation of w_i where w_j is given as context. Equation (3.3) and (3.4) are for intra-cluster and inter-cluster semantic distance measure respectively, *K* is the number of weighted vertices between two lexicons w_i and w_j, v_k is the weighted vertex between two lexicons, *M* is the number of cluster centers between the two lexicons and l_c is the distance between the cluster centers of the two lexicons. The $w_j^{\,p}$ is the polarity of the known given word w_j .

This network has been specifically created for handling the unknown words. A bag-of-words method has been adopted for the prediction of semantic orientation of an unknown word. The bag-of-words chain has been formed with most of the known words, syntactically co-located. A Conditional Random Field (CRF)[27] classifier has been trained on the corpus and effectively used in this method. For example, the probable bag-of-words with *X* as the unknown word can be identified as:

9_11-X-Pentagon-USA-Bush

Discuss-Terrorism-X-President

Middle_East-X-Osama

[27] http://crfpp.sourceforge.net/

The CRF classifier has been trained with a simple list of features: co-occurrence distance, ConceptNet similarity scores and known or unknown word tag based on SentiWordNet. With the help of these very simple features the CRF classifier identifies the most probable bag-of-words to predict the semantic orientation of the given unknown word *X*. Once the target bag-of-words has been identified then the following equation (3.5) is used to calculate the polarity of the unknown word *X*. The main problem is that it is nearly impossible to predict polarity for an unknown word. Standard polarity classifiers generally fall back in performance due to the presence of unknown words but the present Syntactic Co-Occurrence Network is very good to handle unknown or new words. The evaluation section (3.7.2.2) presents more empirical results.

Let us consider that the most probable target bag-of-words for the unknown word *X* has been identified as 'Discuss-0.012-Terrorism-0.0-X-0.23-President' where the scores have been extracted from the Concept.

The equation (3.5) is used to identify the polarity of the unknown word *X* :

$$w_x^p = \sum_{i=0}^{n} e_i * \sum_{j=1}^{n} p_i \text{ ----- (3.5)}$$

where e_i is the edge distance extracted from ConceptNet and the p_i is the polarity information of the word in the bag-of-words.

Syntactic co-occurrence network gives reasonable performance increment over the normal linear sentiment lexicon and the Semantic Network Overlap technique. But it has some limitations. It is difficult to frame an appropriate equation to calculate the semantic orientation within the network as it is very high-dimensional. The framed equations (3.3 and 3.4) produce a less distinguishing value for different bag of words. The variable polarity scores for *"High"* under different context (Figure 3.5) can be calculated as:

$$(\textbf{High}, \text{Sensex}) = \frac{0.3+0.3}{2} = 0.3$$

$$(\textbf{High}, \text{Price}) = \frac{0.22+0.35}{2} = 0.29$$

3.7.2.2 Performance of the Syntactic Co-Occurrence Network

Syntactic co-occurrence measure gives a good increment in performance of the polarity classifier over the Semantic Network Overlap technique. The performance of the syntactic co-occurrence measure has been tested on the MPQA corpus with 70.0% accuracy. The same measure has shown as accuracy of 68.0% for Bengali, which is a good increment over the Semantic Network Overlap technique.

It is observed that near about 45% cases of "*Positivity>0 && Negativity>0*" have been resolved and 43% cases of "$|Positivity - Negativity|$>=0.2" has been resolved by the present Syntactic co-occurrence based technique. The scores are relatively higher than the previous Semantic Network Overlap technique.

Publications

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010(a). ***Phrase-level Polarity Identification for Bengali.*** *In International Journal of Computational Linguistics and Applications (IJCLA)*, Vol. 1, No. 1-2, ISSN 0976-0962, Pages 169-182, Jan-Dec 2010.
   http://www.ijcsit.com/docs/Volume%202/vol2issue1/ijcsit2011020107.pdf

2. **Amitava Das** and Sivaji Bandyopadhyay. 2010(h). ***Opinion-Polarity Identification in Bengali***. *In the Proceeding of the 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL 2010)*, KESE 2010, Pages 41-44, Redwood City, California, USA.
   www.amitavadas.com/Pub/ICCPOL_2010.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2012(c). ***Sentimantics: The Conceptual Spaces for Human Cognition and Sentiment***. *In the Proceeding of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, Delhi, India. (**Communicated**).

Chapter 4

Sentiment Structurization

In the previous three chapters, sentiment knowledge acquisition and representation, subjectivity detection and sentiment polarity classification have been discussed. The need of the end user is the driving force behind the sentiment analysis research. Therefore the outcomes of these research endeavors should lead to the development of a real time sentiment analysis system, which will successfully satisfy the need of the end user. Such a sentiment analysis system should be smart enough to aggregate the scattered sentimental information from the various blogs, news article and from written reviews.

Let us have a look at some real life need of the end user. For example, a market surveyor from company A may have a need to find out the changes in public opinion about their product X after release of product Y by another company B. The different aspects of product Y that the public consider better than product X are also points of interest. These aspects could typically be the durability of the product, power options, weight, color and many other issues that depend on the particular product. In another scenario, a voter may be interested to study the change of public opinion about any leader or any public event before and after any election. In this case the aspect could be a social event, economic recession and may be other issues. The end users are not only looking for the binary (positive/negative) sentiment classification but they are more interested in aspectual sentiment analysis. Therefore only sentiment detection and classification is not enough to satisfy the need of the end user. A sentiment analysis system should be capable enough to understand and extract the aspectual sentiments presented in a natural language text. The sentiment analysis research reported in the present work has been extended towards the sentiment structurization or opinion constituent identification.

The philosophical paradox of the sentiment structurization task is to first find out a generic sentiment structure that is portable across domains and languages. Once the generic sentiment structure has been arrived at, the next subtask is to build an appropriate technology to extract the sentiment details from the text in the proposed structure.

In the section 4.1, attempts have been made to answer the philosophical question "**How to define the proper sentiment/opinion structure?**". The section 4.2 tries to find out the answer of the same philosophical question with the light of previous studies. Section 4.3 elaborates the idea of the proposed 5W (Who, What, When, Where and Why) structurization. In section 4.4 several motivations behind the 5W structurization have been described. The resource acquisition processes typically corpus collection and annotation are reported in the section 4.5. The next sections 4.6 and 4.7 describe the feature extraction and the proposed system architecture for automatic 5W extraction respectively.

4.1 Opinion: The Medium between Knowledge and Ignorance

It is very hard to define sentiment or opinion and to identify the regulating or the controlling factors of sentiment. Moreover, no concise set of psychological forces could be defined that really affect the writers' sentiments, i.e., broadly the human sentiment.

Chapter 4 Sentiment Structurization

"*How the mind works is still a mystery. We understand the hardware, but we don't have a clue about the operating system.*"

James Watson (Nobel laureate)

Although the complete theory of human sentiment is still not yet explored but the title of this section, "Opinion: The Medium between Knowledge and Ignorance", a famous saying by the Greek philosopher Plato, gives an abstract idea about the characteristics of sentiment or opinion. Sentiment is a human instinct. Sentiment is not necessarily driven by proper knowledge or rationality but rather it depends on individuals and varies at individual level.

I always prefer hill stations for my holidays whenever my travel guide try to convince me for the world famous beaches like Mauritius or Miami.

In the previous example sentence, the writer or the speaker is well informed about the world famous beaches but still he/she is inclined towards the hill stations. It means the writer or the speaker is ignoring his/her knowledge and is driven only by his/her likings or sentiment. But sentiment is a bi-directional process and the sentiment of the reader or listener is equally important. The following set of example dialogues put the sentiment focus from the listener's or reader's perspective.

Teacher: Students, how you spend your holidays?

Student 1: Nice. I met my grandparents.

Student 2: Excellent. I used to play cricket with my friends.

Student 3: We spent a nice holiday in Kovalam with family.

……..

……

Principal: How your student feeling after holidays?

Teacher: Oh, they are feeling well as they have spent good time.

The previous dialogue is a good example to understand the sentiment from the perspectives of both the writer or the speaker and the reader or the listener. The Teacher was ignoring the granular knowledge of how each and every student spent their holidays and was concentrating only on their sentiment about how they spent their holidays. This becomes clear when the teacher summarizes the whole communication to the Principal.

The sentiment analysis research along with natural language processing techniques attempt to develop systems that can aggregate sentiment information in the form of retrieved documents, textual or visual summaries and sentiment tracking systems. The pertinent question is to identify the knowledge that is

essential to develop this kind of systems as well as to identify the knowledge that should be ignored. The answer lies on the need of the end user. A market analyst from a company may want to know the reasons for disliking of their products by many users. The identified reasons could be helpful for him to fix further marketing strategy. A casual voter may want to know about the general public opinion for leader X and leader Y but may not be interested in the minute details. Such details could be useful for a news reporter.

The previous discussion points to the necessity of a proper sentiment structurization. Identification of a generic sentiment structure is very difficult. A good number of research endeavors could be found in literature on structured sentiment extraction and these are reported in the next section. In the present work, the 5W (Who, What, When, Where and Why) structurization has been proposed for the generic sentiment structure. The 5W method is more generic and also portable across domains. The 5W task seeks to extract the semantic information of nouns in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. The motivations behind the 5W structurization are detailed in the section 4.4.

4.2 What Knowledge to Acquire and What to Ignore and Why?

There are two key issues of the structured sentiment extraction task:

1. How to define the proper sentiment/opinion structure?
2. How to extract the structured sentiment/opinions?

Various structures have been proposed by several researchers and these structures vary in nature due to the domain for which the system has been developed or due to the targeted output of the system. The sentiment analysis systems identify sentiments on targets that are typically objects and focus on their components, attributes and features. An object can be a product, service, individual, organization, event, topic etc. The previous related research works have been reported in the following subsections based on the identification of common attributes, like, Holder, Topic and argumentation.

4.2.1 Sentiment / Opinion Holder

Identification of sentiment / opinion holder has become a separate research sub-discipline nowadays and has been attempted by a numbers of researchers. Sentiment/opinion holder is treated as an opinion source that needs to be identified for further summarization, tracking or question-answering task.

(Kim and Hovy, 2004) have concluded that an analytic definition of opinion is impossible. Thus they have described opinion as a quadruple (Topic, Holder, Claim and Sentiment). The 'Holder' is identified as a very important aspect in the quadruple for further understanding of the *sentiment of opinions*. It has been hypothesized that **PERSON** and **ORGANIZATION** could be the only possible opinion holders and the

named entity tagger from BBN has been used in the setup with further NLP based contextual disambiguation technique.

(Choi et. al., 2005) view the holder identification as a source identification problem and tackle it using sequence tagging and pattern matching techniques simultaneously. The goal of the source identification is to identify direct and indirect sources of opinions, emotions, sentiments and other *private states* that are expressed in the text. Using syntactic, semantic and orthographic lexical features, dependency parse features and opinion recognition features the system has been trained using a linear-chain Conditional Random Field (CRF) to identify the opinion sources. The CRF treats source identification as a sequence tagging task, while the AutoSlog (Riloff, 1996) views the problem as a pattern-matching task. Reported result proves that the combination of the two techniques perform well than either one alone. The main contribution of this research attempt is the hybrid system architecture for holder identification. The system was evaluated with 79.3% precision and 59.5% recall using a head noun matching measure and 81.2% precision and 60.6% recall using an overlap measure. The semantic hierarchy of opinion sources as proposed in the work is shown in the Figure 4.1.

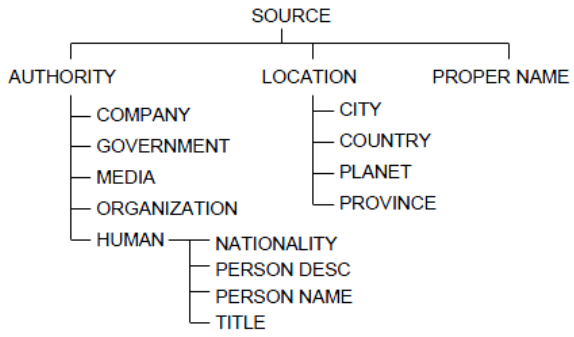


Figure 4.1: The Semantic Hierarchy of Opinion Sources (Choi et. al., 2005)

(Bethard et. al., 2006) have described the automatic extraction of *propositional opinions* mainly from the interest in automatic opinion question answering. Answering an opinion question (e.g., "*How does X feel about Y?*" or "*What do people think about Z?*") requires finding which clauses express the exact opinion of the subject. The key role in opinion question answering is to solve the problem of extracting propositional opinions towards breaking down opinions into their various components. Semantic databases like FrameNet (Baker, Fillmore and Lowe, 1998) and PropBank (Kingsbury, Palmer and Marcus, 2002) in which semantic constituents of sentences are tagged with AGENT, THEME and PROPOSITION information are expected to help in extracting propositional opinions and opinion holders. The extraction of propositional opinion arguments is an extension of the earlier works on semantic parsing (Gildea and Jurafsky, 2002; Pradhan et. al., 2003) with new lexical features that represent opinion words. In the semantic parsing work, sentences were labeled for thematic roles (AGENT, THEME, and PROPOSITION among others) by training statistical classifiers on FrameNet and PropBank. In the propositional opinion extraction technique a modified role label has been used

(OPINION-PROPOSITION versus NULL). Words that are associated with opinions are used as additional features for this model; these words are automatically learned by bootstrapping from smaller sets of such known words. Finally, results are presented from a three-way classification where sentence constituents are labeled as OPINION-PROPOSITION, OPINION-HOLDER or NULL. The most important issue in the work is that it uses existing semantic resources for structured opinion extraction. Unfortunately this technique is not replicable for Bengali, as FrameNet and PropBank are not available.

The system reported in NTCIR-6 (NII Test Collection for IR Systems), by (Bloom et. al., 2007) defines the structured opinion extraction task based on the ***Appraisal Theory*** (Martin and White, 2005). An appraisal expression is an elementary unit of text through which an opinion holder (the source) expresses an opinion (the attitude) about a target. In an appraisal expression, the three text fragments that functions as *source, attitude and target* are connected syntactically and may not be found contiguously in the text. Some of these text functions (e.g., source or target) may not be explicit and may be left by the speaker or writer to be inferred from the context. On a similar note, it has been observed during the experiments that 5W constituents are not present in every sentence. The appraisal theory is a grammatical theory that deals with the representation of opinion in text. The attitude system (Bloom et. al., 2007) classifies evaluative languages into three general types of opinions: ***affect*** (an internal emotional state), ***appreciation*** (of intrinsic qualities of an object) or ***judgment*** (concerning the way people behave). English grammar imposes different constraints on how these three types of appraisals can be expressed. One cannot, for example, talk about "*an evil towel*" very easily because "*evil*" is a type of judgment, but a towel is an object that does not have behaviors (unless anthropomorphized). The reported system is based on a developed general lexicon of words that can be used to express attitudes as well as on a shallow parsing system that finds whole phrases that may carry different sentiment orientation than the single words listed in the lexicon.

4.2.2 Sentiment / Opinion Topic

Sentiment analysis task also involves the target (topic) identification from the opinionated text. (Ku et. al., 2005) present automatic opinion summarization techniques based on topic model. The system selects representative words from a document set to identify the main concepts in the document set. A term is considered to represent a topic if it appears frequently across documents or in each document. Appropriate weights are used at sentence, paragraph or document level to detect the representative topic words. The identified topic is then further used for opinion summarization.

(Yi et. al., 2006) present a sentiment analyzer that extracts sentiment (or opinion) about a subject from online text documents. Instead of classifying the sentiment of an entire document about a subject, the system detects all references to the given subject and determines the sentiment in each of the references using natural language processing (NLP) techniques. The system detects sentiment by topic specific feature term extraction. A sentiment pattern database is used to detect the topic sentiment. The sentiment pattern database contains sentiment extraction patterns for sentence predicates extracted from WordNet *emotion cluster*. Each database entry is defined in the following form:

<predicate> <sent_category> <target>

- predicate: typically a verb
- sent_category: + | - | [˜] source

  source is a sentence component (SP|OP|CP|PP) whose sentiment is transferred to the target. SP, OP, CP and PP represent subject, object, complement (or adjective) and prepositional phrases respectively. The opposite sentiment polarity of source is assigned to the target, if ˜ is specified in front of source.

- target is the sentence component (SP| OP|PP) to which the sentiment is directed.

After automatic detection of sentiment phrases in a sentence, the system searches for matching predicate structure in the sentiment pattern database and tries to identify the topic sentiment. On a similar note, different gazetteers lists have been used in the present work to identify different role labels in the 5W role assignment task.

(Zhou et. al., 2006) have proposed the architecture for summary generation system from blogosphere. Typical multi-document summarization (MDS) systems focus on content selection followed by synthesis that is based on removing redundancy across multiple input documents. The online summarization system works on an online discussion corpus involving multiple participants and discussion topics by various participants. Due to the complex structure of the dialogue, similar subtopic structure identification in the participant-written dialogues is essential. Maximum Entropy Model (MEMM) and Support Vector Machine (SVM) have been used with a number of relevant features for the content selection task.

(Kawai et. al., 2007) have developed a news portal site called Fair News Reader (FNR) that recommends news articles to a user with different sentiments in each of the topics in which the user is interested. The FNR can detect various sentiments of news articles. It can also determine the sentimental preferences of a user based on the sentiments of previously read articles by the same user. News articles crawled from various news sites are stored in a database. The contents are integrated as needed and the summary is presented on one page. A document sentiment vector on the basis of topic word lattice model is generated for every document. The topic words are typically high frequent noun, adjective, adverb and verb words. A weighing mechanism plays a crucial role to identify the topic words from those high frequent words. A user sentiment model has been proposed based on user sentiment state. The user sentiment state model works on the browsing history of the user represented as user sentiment vector. The intersection of the documents under User Sentiment Vector and Document Sentiment Vector identify the different news articles with different sentiments in the requested topics by the user.

(Choi et. al., 2009) have proposed a method for automatic sentiment topic extraction based on domain specific lexical clue set, dependency relations and co-occurrence pattern. The method is based on the hypothesis that a sentiment topic is strongly connected to sentiment clues when a sentiment topic and a

clue share a syntactic dependency in a sentiment-revealing sentence and when topic and sentiment clues co-occur in the domain corpus. Some co-occurrence rules have been developed based on domain specific observation. For example, an adjective clue would be dependent on a noun phrase. The system calculates the score of each sentiment topic candidate based on co-occurrence information and picks the highest ranked candidate as the sentiment topic. During the score calculation process the system first computes the contextual similarity between a noun phrases, i.e., the candidate phrase and the current sentiment clue in the sentence based on the co-occurrence information learnt during training.

The Topic-Sentiment Mixture model proposed by (Mei et. al., 2007) is very effective for any real time topic-sentiment analysis. Figure 4.2 shows a topic-sentiment summary. Given a query word (*Dell Laptop)* that represents the ad-hoc information need of the user, the system extracts the subtopics in the search results and associates each subtopic with positive, negative and neutral sentiment sentences retrieved. The subtopics and the associated sentiment sentences are organized in a two dimensional structure. The topic-sentiment summary helps the user to understand the pros and cons of each aspect of the product.

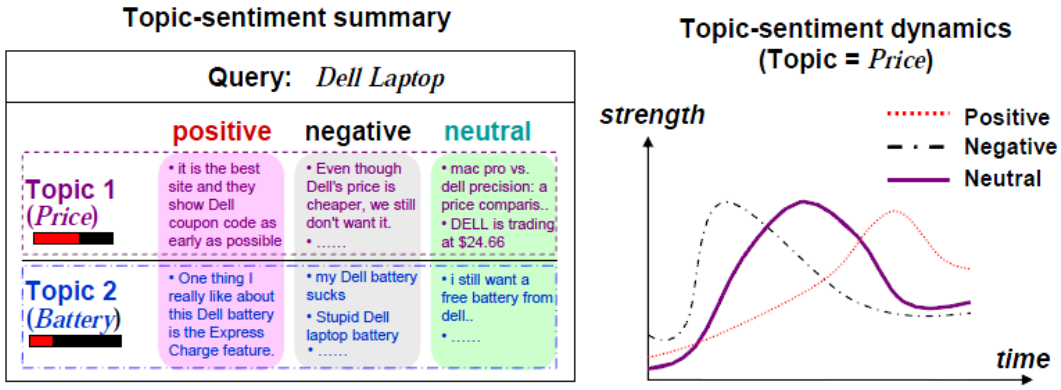


Figure 4.2: A Possible Application of Topic-Sentiment Analysis (Mei et. al., 2007)

4.2.3 What Else?

Few significant contributions could be found in the literature that properly addresses the whole task of opinion structurization. (Kobayashi et. al., 2006) have defined the opinion structure as a frame composed of the following eight constituents:

Opinion holder: A person who is making an evaluation (usually, either the author or an unspecified person)

Subject: A named entity (product or company) of a given particular class of interest (e.g., a car model name in the automobile domain).

Part: A part, member or related object of the subject with respect to which the evaluation is made (engine, interior etc. in the automobile domain)

Attribute: An attribute (of a part) of the subject with respect to which the evaluation is made (size, color, design etc.)

Evaluation: An evaluative or subjective phrase used to express an evaluation or the opinion holder's mental/emotional attitude (good, poor, powerful, stylish, (I) like, (I) am satisfied etc.)

Condition: A condition under which the evaluation applies (driving on winding roads, when traveling with a family etc.)

Support: An objective fact or experience described as a supporting factor of the evaluation (weighs nearly 1,500 kg. etc.)

It was hypothesized that these eight constituents can best depict all the semantic dimensions of an opinion. But, in reality many difficulties may be faced to portrait opinions from different domains as product reviews, movie reviews, social opinions etc. into these eight constituents. Constituents like part and attribute are very much integral parts of domain typology, whereas, opinions from general domain such as social issues cannot be categorized in the same manner. Opinions from social domains generally reflect several events or sub events, which cannot be categorized as part and attribute. In the present work, the proposed constituent category "*what*" is a very good generalization and can portrait all aspects in any domain. The defined typology is a generalization and can be best described as "Opinion on what?".

Other important opinion constituents are 'condition' and 'support'. These identified opinion constituents may be used for further task such as opinion argumentation (Bal and Saint-Dizier, 2009). Instead of identifying so many categories at the basic level, the category has been generalized as "why". The 'why' category may be further analyzed by an opinion argumentation analysis system.

Two important dimensions are missing in the formal opinion constituent's definition by (Kobayashi et. al., 2006). Temporal and locative markers are important features that should be identified for further task like opinion tracking (Ku et. al., 2006). In the present task, these have been defined as the "When" and "Where" constituents of opinion.

In general any kind of tagging scheme depends on the output characteristics of any system. Therefore a structured opinion tagging scheme should depend on the output category but till now, no generalized output characteristics of sentiment / opinion analysis system could be defined. (Dasgupta and Ng, 2009) have thrown an important question: "***Topic-wise, Sentiment-wise, or Otherwise?***" about the output category of the proposed sentiment analysis systems. It really depends on the end user requirement as well as the domain characteristics. So, domain adaptation is definitely an important issue for any system but the goal of the present work is to generate a baseline opinion structure that works across domains and languages. According to the best of our knowledge, no such system could be found in the literature on generalized structure for opinion extraction. The present identification and categorization

task for the semantic constituents of opinion is modeled as 5Ws distilling technique. The technique follows a globally simplified general architecture and is useful for any domain and language.

4.3 The Proposed 5W Rationale

For sentiment/opinion structurization the 5W constituent extraction technique has been proposed (Das et. al, 2010(i)). The developed system identifies and extracts structured opinions by means of semantic constituents using 5W distilling in each document. The 5W task seeks to extract the semantic information of opinion constituents in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. The 5W questions are as follow:

- Who? Who was involved?
- What? What happened?
- When? When did it take place?
- Where? Where did it take place?
- Why? Why did it happen?

The ideas of 5Ws have been used successfully for a machine translation evaluation methodology (Kristen et al, 2009). The methodology addresses the cross-lingual 5W task: given a source language sentence and the corresponding target language sentence, it evaluates whether the 5Ws in the source have been comprehensibly translated into the target language.

The proposed 5W structure is domain independent and more generic than the existing semantic constituent extraction structure. In the next section a comparative and evaluative description of the 5W concept is presented with the contemporary and historical semantic role theories.

4.4 The Motivations behind the 5W Concept

In journalism, the five Ws (Who, What, When, Where and Why) is a concept in news style, research and police investigations that are regarded as the basics in information gathering. The concept of 5Ws was first introduced by (Philip, 1949) in journalism. There is a close similarity of the 5W concept with the Paninian karaka theory and the Fillmore's case grammar.

4.4.1 Panini's Karaka Theory

The study of semantic roles started since Panini's karaka theory that assigns generic semantic roles to words in a natural language sentence. The classical Sanskrit grammar Astadhyayi[1] ('Eight Books'), created by the Indian grammarian Panini at a time variously estimated at 600 or 300 B.C. (Robins, 1979), includes a sophisticated theory of thematic structure that remains influential till today. Panini's Sanskrit grammar is a system of rules for converting semantic representation of sentences into phonetic representations (Paul and Staal, 1969). This derivation proceeds through two intermediate stages: the level of karaka relations, which are comparable to the thematic role types; and the level of morpho-syntax. The proposed grammar rules by Panini map each of the karakas to a basic semantic relation and a basic morpho-syntactic expression. More specialized variants of both types of rules are specified as well, with the basic relation and basic expression acting as defaults whenever the conditions for the variants are not met.

For example, the basic semantic relation of the **apadana** karaka (Source: Where/When) is defined on the fixed point from which something recedes. But with certain verbs **apadana** is used for special relations such as the source of fear, the object someone is hiding, hindering or learning from and so on. The basic expression of **apadana** karaka is Ablative case. The basic semantic relation of the **karma** karaka (Theme: What) refers to the object that is primarily desired; its basic expression is Accusative case. The basic semantic relation of **karana** karaka (Instrument: What) refers to the most effective means of executing the action. While its basic expression is Instrumental case, some verbs are instead specified for the Genitive case to express the **karana** (such as 'break', 'eat', etc.) karaka. Other karakas include **sampradana** (Indirect Object: What), **adhikarana** (Locative: Where), **karta** (Agent: Who) and **hetu** (Cause: Why).

The present work aims for easy implementation of Panini's karaka theory at the crossroads of syntactic to semantic formalization of language aspects. However, on a closer look, several complications arise, especially in Panini's recourse to semantics in many of the ***vidhi*** or ***samajhna*** rules. This seems to happen more in the ***karaka prakarana*** than in other components. An important effort (Vaidya et. al., 2009) describes a syntactic annotation scheme for English based on Panini's concept of karakas.

The present work focuses on the semantic aspects of Panini's karaka theory using the simple and robust 5W distilling process and highlights the challenges in implementing these rules. Standard semantic role labels are not used in the present work as simple 5W concepts can be easily mapped to Panini's karaka theory to robustly describe the syntactic and semantic synergy of any natural language.

4.4.2 Semantic Roles in Modern Generative Grammar

Fillmore's Case Grammar (Fillmore, 1968) revived Panini's proposals in a modern setting. The main objective of Case Grammar was to identify semantic argument positions that may have different realizations in syntax. Fillmore hypothesized "*a set of universal, presumably innate, concepts which*

[1] http://en.wikipedia.org/wiki/P%C4%81%E1%B9%87ini

identify certain types of judgments human beings are capable of making about the events that are going on around them". He posited the following preliminary list of cases, noting however that 'additional cases will surely be needed' (and indeed Fillmore added more in later works (Fillmore et. al., 2003)).

- **Agent:** The typically animate perceived instigator of the action. (**Who**)
- **Instrument:** Inanimate force or object causally involved in the action or state. (**What**)
- **Dative:** The animate being affected by the state or action. (**Who**)
- **Factitive:** The object or being resulting from the action or state. (**What**)
- **Locative:** The location or time-spatial orientation of the state or action. (**Where/When**)
- **Objective:** The semantically most neutral case conceivably the concept should be limited to things which are affected by the action or state. (**Why**)

4.4.3 Recent Trends of Semantic Role Labeling

In the last few years there has been an increased interest in shallow semantic parsing of natural languages as an important component in all kinds of Natural Language Processing (NLP) applications. Semantic Role Labeling (SRL) is a shallow semantic parsing technique that is now being widely used in question and answering (QA), information retrieval (IR) and information extraction (IE), machine translation, paraphrasing, textual entailment, event tracking and so on. The SRL task is to assign semantic roles of predicates (most frequently verbs) at sentence level to syntactic constituents (arguments). A semantic role is the relationship that a syntactic constituent has with a predicate. Given a sentence, the task consists of analyzing the propositions expressed by some target verbs of the sentence. In particular, for each target verb all the constituents in the sentence have to be recognized that fill a semantic role of the verb. Typical semantic arguments include Agent, Patient, Instrument, etc. and also adjuncts such as Locative, Temporal, Manner, Cause, etc.

SRL has been extensively studied for English language but no such effort could be found in Indian languages and especially in Bengali. A linguistic annotation task for Hindi SRL is reported in (Palmer et. al., 2009). The present work reports the development of resources and methodologies to extract semantic role labels of Bengali nouns using 5W distilling.

Semantic roles are generally domain specific in nature such as FROM_DESTINATION, TO_DESTINATION, DEPARTURE_TIME etc. Verb-specific semantic roles have also been defined such as EATER and EATEN for the verb *eat*. The standard lexical resource that is widely used in various English SRL systems is PropBank (Palmer et. al., 2005; Fillmore et. al., 2003; Kipper et. al., 2006). These collections contain manually developed well-trusted gold reference annotations of both syntactic and predicate-argument structures.

PropBank defines semantic roles for each verb. The various semantic roles identified (Dowty, 1991) are Agent, patient or theme etc. In addition to verb-specific roles, PropBank defines several other general roles that can apply to any verb (Palmer et. al., 2005).

FrameNet[2] is annotated with verb frame semantics and is supported by corpus evidence. The frame-to-frame relations defined in FrameNet are Inheritance, Perspective_on, Subframe, Precedes, Inchoative_of, Causative_of and Using. Frame development focuses on paraphrasability (or near paraphrasability) of words and multi-words.

VerbNet annotated with thematic roles refer to the underlying semantic relationship between a predicate and its arguments. The semantic tagset of VerbNet consists of tags such as agent, patient, theme, experiencer, stimulus, instrument, location, source, goal, recipient, benefactive etc.

It is evident from the above discussions that no adequate semantic role set exists that can be defined across various domains. Researchers generally rely on a customized tagset, developed on the basis of the necessity of the particular nature of any problem. As no concise set of semantic role labels exist, the development of a generic semantic tagset that will be portable across domains and languages has been particularly concentrated in the present work. The idea has been explored for Bengali language.

From the next section, the technical challenges faced during the development of the automatic 5W extraction system have been discussed. In the next chapter, answer to the question "*Does the generic 5W structure remain useful in the real life systems?*" has been attempted. It has been shown that the 5W constituents are effective for the sentiment summarization and tracking.

4.5 Resource Organization

Resource acquisition is one of the most challenging tasks while working with resource constrained languages like Bengali. Bengali is the fifth popular language in the World, second in India and the national language of Bangladesh. Extensive NLP research activities in Bengali have started recently but resources like annotated corpus, various linguistic tools are still unavailable for Bengali in the required measure. In the present work, the manual annotation of the gold standard Bengali corpus has been attempted.

4.5.1 Corpus

For the present task, the corpus from the ICON 2009 Dependency Parsing shared task[3] has been chosen. The data is manually annotated with part of speech (POS), chunk, morphological features and dependency tree relationships. Detailed reports about this corpus development in Bengali can be found in (Ghosh et. al., 2009). The corpus statistics is presented in Table 4.1.

[2] https://framenet.icsi.berkeley.edu/fndrupal/
[3] http://ltrc.iiit.ac.in/nlptools2009/

Bengali Corpus Statistics			
	Training	Development	Test
Total number of sentences in the corpus	980	150	150
Total number of wordforms in the corpus	9223	1762	1812
Total number of distinct wordforms in the corpus	6233	522	628

Table 4.1: Statistics of 5W Annotated Bengali News Corpus

4.5.2 Annotation

Sanchay[4], a well known linguistic annotation tool for Indian languages has been used for Bengali sentence level 5Ws manual annotation task. Two annotators (Mr. X and Mr. Y) participated in the present task. The annotated documents are saved in Shakti Standard Format[5] (SSF: XML format).

Annotators were asked to annotate 5Ws in Bengali sentences in terms of Bengali chunks. Instructions have been given to annotators to find out the main finite verb in a sentence and successively extract 5W components by asking 5W questions to the main verb. The annotators summarize the information in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. An example of the 5Ws annotated document is presented in Figure 4.3. The chunk heads (e.g., 1 ((NP <fs af='মাধবীলতা,n,,sg,,d,0,0' head="মাধবীলতা", **Who**>) contain the W tag in last position of the feature structure (fs).

```
<Sentence id="1">
1       ((      NP      <fs af='মাধবীলতা,n,,sg,,d,0,0' head="মাধবীলতা", Who>
1.1     মাধবীলতা       NN      <fs af='মাধবীলতা,n,,sg,,d,0,0' name="মাধবীলতা">
        ))
2       ((      VGNF    <fs af='শো,v,,,2,,বে,be' head="শোবে", Why>
2.1     শোবে    VM      <fs af='শো,v,,,2,,বে,be' name="শোবে">
        ))
3       ((      VGF     <fs af='বল্,v,,,3,,নে,ne' head="বলে">
3.1     বলে     VM      <fs af='বল্,v,,,3,,নে,ne' name="বলে">
        ))
4       ((      NP      <fs af='তখন,pn,,,,d,0,0' head="তখন", When>
4.1     তখন     PRP     <fs af='তখন,pn,,,,d,0,0' name="তখন">
        ))
5       ((      NP      <fs af='হাত,n,,sg,,o,এর,era' head="হাতের">
5.1     হাতের   NN      <fs af='হাত,n,,sg,,o,এর,era' name="হাতের">
```

[4] http://sourceforge.net/projects/nlp-sanchay/
[5] http://web2py.iiit.ac.in/publications/default/view_publication/techreport/54

```
        ))
6       ((      NP      <fs af='ঘড়ি,unk,,,,,,' head="ঘড়ি" poslcat="NM", What>
6.1     ঘড়ি    NN      <fs af='ঘড়ি,unk,,,,,,' name="ঘড়ি" poslcat="NM">
        ))
7       ((      VGNF    <fs af='খুল্,v,,,3,,নে,ne' head="খুলে">
7.1     খুলে    VM      <fs af='খুল্,v,,,3,,নে,ne' name="খুলে">
        ))
8       ((      NP      <fs af='টেবিল,n,,sg,,d,মে,me' head="টেবিলে", Where>
8.1     টেবিলে  NN      <fs af='টেবিল,n,,sg,,d,মে,me' name="টেবিলে">
        ))
9       ((      VGF     <fs af='রাখ্,v,,,3,,ছিল,Cila' head="রাখছিল">
9.1     রাখছিল  VM      <fs af='রাখ্,v,,,3,,ছিল,Cila' name="রাখছিল">
9.2     SYM     <fs af='.,punc,,,,,,' poslcat="NM">
        ))
</Sentence>
```

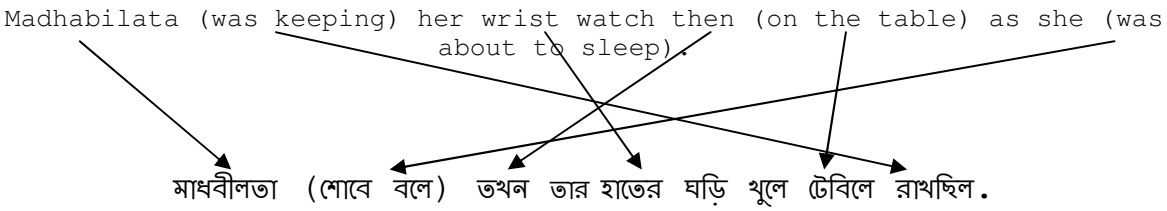


Figure 4.3: An Example of Chunk Level 5W Annotated Sentence

4.5.3 Inter-Annotator Agreement

The agreement of annotations between two annotators has been evaluated. The agreements of tag values at each 5W level are listed in Table 4.2.

Tag	Annotators X and Y Agreement percentage
Who	88.45%
What	64.66%
When	76.45%
Where	75.23%
Why	56.23%

Table 4.2: Inter-Annotator Agreement at Each W Level

It has been observed that in the present task the inter-annotator agreement is better for Who, When and Where level annotation rather than for What and Why level annotation. It is difficult to reach a conclusion at this point since only a small number of documents have been considered.

Further discussion with annotators reveals that the psychology of annotators is to grasp all 5Ws in every sentence, whereas practically all 5Ws are not present in every sentence. It is observed from Table 4.2 that most ambiguous tag is “What”. Let us consider an example.

রাম/Who শ্যামের কানে/Where কানে/Where বললো।

Ram/Who whispers at Shyam’s ear/Where.

In the preceding example Ram/রাম should be tagged as “Who” but the identification of the candidate for “What” tag is ambiguous. One annotator tagged Shyam’s/শ্যামের as “What”, but it is an animate object of the main verb whisper/বললো. Conceptually animate objects should be categorized as “Whom”. In that case 5Ws should be listed as Who, What/Whom, When, Where and Why or 6Ws that include “Whom”.

It is shown in Table 4.3 that the co-occurrence of “What” and “Who” tag is 58.56% in the overall corpus. That means “Who” and “What” have occurred 58.56% in the same sentence based on the total number of sentences in the corpus. There is a good number of cases where “What” refers to an animate object. But in the present task and for the sake of simplicity only the inanimate objects have been considered for annotating with “What” tags.

Tags	Percentage					
	Who	What	When	Where	Why	Overall
Who	-	58.56%	73.34%	78.01%	28.33%	73.50%
What	58.56%	-	62.89%	70.63%	64.91%	64.23%
When	73.34%	62.89%	-	48.63%	23.66%	57.23%
Where	78.0%	70.63%	48.63%	-	12.02%	68.65%
Why	28.33%	64.91%	23.66%	12.02%	-	32.00%

Table 4.3: Sentence Level Co-occurrence Pattern of 5Ws

It has also been observed that 5W annotation task takes very little time for annotation with only a small number of clearly defined tags. Annotation is an important yet tedious task for any new data driven experiment in NLP, but 5W annotation task is easy to adopt for any new language.

4.6 Feature Extraction

The effective set of features is to be found through a series of experiments. Bengali is an electronically resource scarce language in terms of NLP tasks. The aim in the present task is to find the minimum yet effective set of features. More number of features means more NLP tools, which may not be readily available for the language. All the features that have been used to develop the present system are categorized as Lexical, Morphological and Syntactic features. These are listed in the Table 4.4 below and

have been described in the subsequent subsections. The Bengali Shallow Parser[6] developed under Indian Languages to Indian Languages Machine Translation (IL-ILMT) project has been used in the present work.

Types	Features	
Lexical	POS	
	Root Word	
Morphological	**Noun**	Gender
		Number
		Person
		Case
	Verb	Voice
		Modality
Syntactic	Head Noun	
	Chunk Type	
	Dependency Relations	

Table 4.4: Features for 5W Role Labeling Task

4.6.1 Lexical Features

Lexical features are the basic linguistic clues to identify the semantic role of any predicate. The following two features have been used in the present system.

4.6.1.1 Part of Speech (POS)

POS of any word cannot be treated as a direct clue of its semantics but it definitely helps to identify it. Finding out the POS of any word can reduce the search space for semantic meaning. It has been shown by (Gildea and Jurafsky, 2002), (Palmer et. al., 2005) etc. that the part of speech of any word in sentences is a vital clue to identify the semantic role of that word. The Bengali Shallow parser extracts the POS of a word.

4.6.1.2 Root Word

Root word is a good feature to identify word level semantic role especially for those types of 5Ws like "When", "Where" and "Why" where manual dictionaries have been made. There are various conjuncts and postpositions, which directly indicate the type of predicate present in any sentence. For example, জন্য, হেতু give clue that the next predicate is causative ("Why"). The Bengali Shallow parser extracts the root of a word.

[6] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

4.6.2 Morphological Features

The noun phrases can only hold the 5W semantic roles. Therefore, morphological features of noun are very essential. But the semantic roles depend on the relation of these noun phrases with the principal verb in the same sentence. Thus morphological features of the verbs are also taken into consideration.

4.6.2.1 Nouns

4.6.2.1.1 GENDER

Gender information of a noun chunk is essential to relate the chunk to the principle verb modality. In the case of "What"/"Whom" ambiguities gender information of the noun word significantly helps the disambiguation. The gender information is null for inanimate objects while animates have a definite value. Bengali is not a gender sensitive language, i.e., the verb form does not vary on the gender of the subject or the adjective form does not depend on the gender of the accompanying noun word. The gender feature is not significant rather than the number and person features. But the statistical co-occurrence of gender information with the number and person information is significant.

4.6.2.1.2 NUMBER

Number information of a noun chunk helps to disambiguate the "Who"/"What" ambiguities. In the section 4.5.3 on inter-annotator agreement, "Who" has been identified first by matching the modality information of the principle verb with the corresponding number information of the noun chunk.

4.6.2.1.3 PERSON

Person information of a noun chunk is as important as the number information. It helps to relate the head of any noun chunk to the principle verb in a sentence.

4.6.2.1.4 CASE

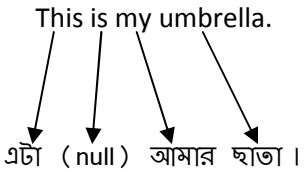
Case markers are generally described as the *karaka* relation of a noun chunk with the main verb. Semantically, *karaka* is the ancestor of all semantic role interpretations. Case markers are categorized as Nominative, Accusative, Genitive and Locative. Case markers are very helpful for most of the 5W semantic role identification tasks.

4.6.2.2 Verb

4.6.2.2.1 Voice

The distinction between active and passive verbs plays an important role in relating the semantic roles and grammatical functions, since direct objects of active verbs often play the same semantic roles as the subjects of passive verbs (Gildea and Jurafsky, 2002; Park and Rim, 2005; Pradhan et. al., 2003). A set of hand-written rules helps to identify the voice of any verb chunk. The rules rely on the presence of auxiliary verbs like হয়েছে, হোক etc. which indicates that the main verb in that particular verb chunk is in

passive form. In Bengali, active voice sentences generally drop the copula as in the following sentence the Bengali verb হয় (English equivalent 'is') is dropped. Rules are defined from a standard Bengali grammar book.



4.6.2.2.2 Modality

Honorific markers are very distinctly used in Bengali and it is directly reflected by the modality marker of the verb. For example, the honorific variations of 'করা/do' are কর (used with তুই: 2nd person either of same age or younger), করো (used with তুমি: 2nd person either of same age or slightly elder) and করুন (used with আপনি: 2nd persond generally for aged or honorable person). The verb modality information helps to identify the "Who" tag. "Who" is identified first by matching modality information of principle verb with corresponding number information of the noun chunk.

4.6.3 Syntactic Features

The following syntactic features are used in the present system.

4.6.3.1 Head Noun

The present 5W SRL system identifies chunk level semantic roles. Morphological features of the chunk head are more important rather than the features associated with other chunk members. Head words of noun phrases can identify the selectional restrictions on the semantic role types of the noun chunks. For example, in a communication frame, noun phrases headed by *Ram, brother or he* are more likely to be the SPEAKER (Who), while those headed by proposal, story or question are more likely to be the TOPIC (What).

4.6.3.2 Chunk Type

The present 5W SRL system identifies the semantic roles at the level of noun chunks. Hence chunk level information is an effective feature in the supervised classifier as well as in rule-based post processor.

4.6.3.3 Dependency Relations

It has been established that dependency phrase-structures are most crucial to understand the semantic contribution of a syntactic phrase in a sentence (Gildea and Jurafsky, 2002; Palmer et. al., 2005). A statistical dependency parser has been used for Bengali (Ghosh et. al., 2009).

4.7 Semantic Roles Identification

The 5Ws semantic role labeling task addresses the following NLP issues: predicate identification, argument extraction, attachment disambiguation, location and time expression recognition. To solve these issues the system has been developed on hybrid architecture, Machine Learning technique followed by a rule-based methodology.

One of the most important milestones in SRL literature is CoNLL-2005 Shared Task on Semantic Role Labeling[7]. System reports of those participated systems show that Maximum Entropy (MEMM) based models work well in this problem domain as 8 among 19 systems used MEMM as the solution architecture. The second best performing system (Haghighi et. al., 2005) uses MEMM model based on syntactic information without any pre or post processing.

Table 4.3 presents the sentence level co-occurrence patterns of the 5Ws in the corpus. It is very clear that the co-occurrence pattern of 5Ws is not very regular in the corpus. Hence sequence labeling with 5Ws tags using MEMM will lead to a label bias problem and may not be an acceptable solution for present problem of 5W role labeling (Haghighi et. al., 2005). The proposed system follows a hybrid architecture that statistically (based on MEMM models) assigns 5W labels to each chunk in a sentence and a rule based post-processor helps to reduce many false positives by the MEMM based system and at the same time identifies new 5W labels. This increases the overall performance of the final system. The rule based post-processor works on the output of the MEMM based system. The rules have been developed based on the acquired statistics on the training set and the linguistic analysis of standard Bengali grammar.

4.7.1 Using Maximum Entropy Model (MEMM)

MEMM[8] treats 5Ws semantic role labeling task as a sequence tagging task. The final set of features (described in Table 4.4) has been identified based on feature engineering. All features are binary for the MEMM model. The performance of the 5W SRL task by MEMM is reported in Table 4.8.

It is observed that the performance of the MEMM-based model differs for the various tags. While precision values for “Who”, “When” and “Where” is good but recall is not good as the system has failed to identify these tags in various cases. For the “What” tag, the system identifies most of the occurrences and hence the recall is high. But there are many false hits keeping the precision low. For the “Why” tag, both precision and recall values are low.

4.7.2 Rule-Based Post-Processing

The rules for each tag label are described in the next sub-sections.

[7] http://www.lsi.upc.edu/~srlconll/
[8] http://maxent.sourceforge.net/

4.7.2.1 Who? Who was involved?

The system fails to identify "Who" in many cases. Let us consider the following example sentence:

নিমন্ত্রিত না হলেও তোমার/**Who** যাওয়া উচিত ছিল সেখানে।

Though you are not invited but you/**Who** should go there.

The system fails in this case, because the targeted chunk head is a pronoun and it is placed in the middle of the sentence whereas words with "Who" tags are generally placed at sentence initial positions as Bengali is a Verb final and Subject initial language. The system makes some false hits also. Let us consider the following example sentence:

দরজাটা বন্ধ করো।

Close the door.

In the sentence the word দরজাটা/the door is marked as "Who" whereas the "Who" tag should have been associated with the elliptical "you" (2nd person singular number) in the sentence. The MEMM based system is quite biased towards the chunks at the initial position of sentences.

Rules have been developed using case marker, Gender-Number-Person (GNP), subject and verb modality features to identify words for 'who' tags. These rules increase the overall system performance value as reported in Table 4.8.

4.7.2.2 What? What happened?

As described in the sections 4.5.3, "What" can also be described as "Whom" when the object is animate. To avoid further ambiguities, both animate and inanimate objects have been categorized as "What" for the present task. The corpus distribution of "What" and "Whom" is observed as almost 50%-50%. In the first example sentence, the word বাঁশি/Flute semantically represents the "What" category whereas in the second sentence the word তাকে/him semantically represents the "Whom" category though in the present task it will be tagged as "What" category.

শ্যামের বাঁশি।

Flute of Shyam.

এটা তাকে দিও।

Give this to him.

The positional information of a word is used for "What" or object identification. There is minimal syntactic, orthographic and morphological difference between "Who" and "What". In the present task,

candidates for “Who” are first detected and “What” tag is assigned to the rest of the noun chunks based on the position in the sentence. Significant increment in result due to the application of the appropriate rules has been reported in Table 4.8.

4.7.2.3 When? When did it take place?

Identification of time expressions has an important aspect in NLP applications. People generally study time expressions to track events or in information retrieval tasks. Time expressions have been incorporated in the SRL task with “when” tags.

Time expressions can be categorized into the two types, General and Relative, as listed in Table 4.5. This manually augmented list with pre defined categories is used in the rule based post-processor. There are still many difficulties to identify special cases of relative time expressions. Let us consider the following example sentence:

চাঁদ উঠলে আমরা রওনা হবো।

When the moon rises we will start our journey.

	Bengali		English Gloss
General	সকাল/সন্ধ্যে/রাত/ভোর…		Morning/evening/night/dawn…
	_টার সময়/সময়/ঘটিকায়/মিনিট/সেকেন্ড…		O clock/time/hour/minute/second…
	সোমবার/মঙ্গলবার/রবিবার…		Monday/Tuesday/Sunday…
	বৈশাখ/জৈষ্ঠ/…		Bengali months…
	জানুয়ারী/ফেব্রুয়ারী…		January/February…
	দিন/মাস/বছর…		Day/month/year…
	কাল/ক্ষন/পল…		Long time/moment…
Relative	আগে/পরে…		Before/After…
	সামনে/পেছনে…		Upcoming/
	Special Cases	উঠলে/থামলে…	When rise/When stop…

Table 4.5: Categories of Time Expressions

The relative time expression is উঠলে/**when** rise which is tagged as infinite verb (for Bengali, tag label is VGNF). The present system considers only nouns for tagging of semantic role labels. Hence, words with verb tags are not considered. It has been observed that the occurrence of verb words as relative time expressions is approximately 1.8-2% in the overall corpus.

Similar to "who" tags, the use of the manually augmented list of time expressions followed by application of some hand crafted rules increase the identification of "When" tags in Bengali sentences. The performance increase in terms of recall value is reported in Table 4.8.

4.7.2.4 Where? Where did it take place?

Identification of "Where" semantic role labels simply refer to the task of identification of locative markers in NLP. Locative markers can be categorized into two types, general and relative, and are listed in Table 4.6. This manually edited list is used with the rule based system. Morphological locative case marker feature has been successfully used in the identification of locative markers. There is an ambiguity among "Who", "When" and "Where" tag as they orthographically generate the same type of surface form (using common suffixes as: ে, ের etc). Moreover, it has been observed that there are fewer differences among their syntactic dependency structures throughout the corpus.

দেশে কাজ নেই বাবু।

There is unemployment in the country side.

	Bengali	English Gloss
General	মাঠে/ঘাটে/রাস্তায়	Morning/evening/night/dawn…
Relative	আগে/পরে…	Before/After…
	সামনে/পেছনে…	Front/Behind

Table 4.6: Categories of Locative Expressions

The machine learning based system assigns the "Who" tag to দেশে/country_side as it is present in the initial position of the sentence. Hence rules have been formulated using only morphological locative marker.

A different type of problem has been observed where a verb word plays the "Where" semantic role. Let us consider the example.

লোকে যেখানে কাজ করে সেখানে।

Where people works there.

Here যেখানে কাজ করে/*Where people works* should be tagged as "Where". But this is a verb chunk and the present work considers noun words within its scope. Significant change in performance due to the application of rules is reported in Table 4.8.

4.7.2.5 Why? Why did it happen?

The semantic role assignment for "Why" is the most challenging task. The task is also known as the argument identification. As reported in previously in inter-annotator agreement section (4.5.3), the overall distribution regularity is very low for the role "Why". Irregular and small occurrences of "Why" lead to poor result in the ML-based techniques. Inter-annotator agreement shows that even human annotators disagree on the "Why" tag. To resolve this problem, a relatively large corpus is required to learn fruitful feature similarities among argument structures.

A manually generated list of causative postpositional words and pair wise conjuncts has been prepared to identify argument phrases in Bengali sentences. A snapshot of this list is reported in Table 4.7. Small incremental changes have been noticed in the precision value of "Why" tag identification but no significant increase in recall has been noticed as reported in Table 4.8.

General	Bengali	English Gloss
	জন্য/কারনে/হেতু…	Hence/Reason/Reason
Relative	যদি_তবে	If_else
	যদিও_তবুও	If_else

Table 4.7: Categories of Causative Expressions

4.7.3 Performance of 5W Role Labeling by MEMM and Rule-Based Post Processing

The performance of the 5W Role labeling task using MEMM machine learning algorithm followed by the application of rule base post-processing techniques has been reported in Table 4.8.

Tag	Precision		Recall		F-measure		Avg. F-Measure	
	1	2	1	2	1	2	1	2
Who	76.23%	79.56%	64.33%	72.62%	69.77%	75.93%	62.22%	68.10%
What	61.23%	65.45%	51.34%	59.64%	55.85%	62.41%		
When	69.23%	73.35%	58.56%	65.96%	63.44%	69.45%		
Where	70.01%	77.66%	60.00%	69.66%	64.61%	73.44%		
Why	76.23%	63.50%	53.87%	55.56%	57.41%	59.26%		

Table 4.8: Performance of 5W Role Labeling by MEMM + Rule-Based Post Processing

The precision, recall and the F-measure values for the various semantic role labels using the machine learning technique have been listed in the columns marked (1). After the corresponding rule based post processor has been applied for each semantic role label, the values of the various evaluation metrics are

listed in the columns marked (2). It is observed that the application of the rule based post processing systems always increase the precision, recall and the F-measure values. The average F-measure for all the semantic role labels increases from 62.22% with MEMM based system to 68.10% when rule based post processing is applied.

Publications

1. **Amitava Das**, Aniruddha Ghosh and Sivaji Bandyopadhyay. 2010. ***Semantic Role Labeling for Bengali Noun using 5Ws: Who, What, When, Where and Why***. *In the Proceeding of the International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLPKE2010)*, Pages 1-8, Beijing, China, 2010.
   http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5587772

Chapter 5

Sentiment Summarization-Visualization-Tracking

So far, the analysis and extraction of sentimental/opinionated information from electronic text have been discussed. The focus of this chapter is on aggregating and representing sentiment information drawn from an individual document or from a collection of documents. Sentiment/opinion aggregation is a necessary requirement at the end users' point. For example, an end user might want to have an at-a-glance presentation of the main points made in a single review or how opinion changes from time to time over multiple documents. An intelligent system should be smart enough to aggregate all the scattered sentimental information on any specific topic from the various blogs, news article and from written reviews. The role of such a system is to minimize the human effort and to produce a good sensible output.

There is no doubt that aggregation of sentiment is necessary but it is very hard to achieve a consensus among researchers on how the sentimental information should be aggregated. Although a few systems like Twitter Sentiment Analysis Tool[1], TweetFeel[2] are available in the World Wide Web since last few years still more research efforts are necessary to match the user satisfaction level and social need. This issue has been addressed in the section 5.1 in the light of previous works. To meet the end users requirement the concept of 5W Sentiment Summarization-Visualization-Tracking has been introduced in the section 5.2. Experiments have been started with multi-document topic-opinion summarization and finally the 5W constituent based summarization with visualization and tracking system has been developed. The details of the developed Multi-Document Topic-Opinion summarization and 5W Summarization-Visualization-Tracking systems are described in the sections 5.3 and 5.4 respectively.

5.1 What Previous Studies Suggest, Opinion Summary: Topic-Wise, Polarity-Wise or Other-Wise?

Aggregation of information is the necessity from the end users' perspective but it is nearly impossible to develop consensus on the output format or how the data should be aggregated. Researchers have tried with various types of output format like textual or visual summary or overall tracking with time dimension. The next key issue is "*how the data should be aggregated?*". Dasgupta and Ng (Dasgupta and Ng, 2009) throw an important question: "***Topic-wise, Sentiment-wise, or Otherwise?***" about the opinion summary generation techniques. Actually the output format varies on end users' requirements and the domain. Instead of digging for the answer on the possible output format, experiments have been carried out with multiple output formats. Initially, the topic-wise, polarity-wise and other-wise summarization systems proposed by various researchers have been looked into in the following subsections.

5.1.1 Topic-Wise

There is clearly a tight connection between extraction of topic-based information from a single document and topic-based summarization of that document, since the information that is pulled out

[1] http://twittersentiment.appspot.com/
[2] http://www.tweetfeel.com/

can serve as a summary. Obviously, this connection between extraction and summarization holds in the case of sentiment-based summarization, as well.

(Yi et. al., 2003) present a sentiment analyzer that extracts sentiment (or opinion) about a subject from online text documents. Instead of classifying the sentiment of an entire document about a subject, the system detects all references to the given subject, and determines sentiment in each of the references using natural language processing (NLP) techniques. The system detects sentiment by topic specific feature term extraction. The authors have used a sentiment pattern database (mostly syntactic, e.g., JJ NN or ADV VB etc.) to detect the topic-sentiment. The sentiment pattern database contains sentiment extraction patterns for sentence predicates extracted from WordNet *emotion cluster*. The database entry is defined in the following form:

`<predicate> <sent_category> <target>`

- `predicate`: typically a verb (mostly finite principal verb phase)
- `sent_category`: sentiment category + | - | [~] source

  `source` is a sentence component `(SP|OP|CP|PP)` whose sentiment is transferred to the target. `SP, OP, CP and PP` represent subject, object, complement (or adjective), and prepositional phrases, respectively. The opposite sentiment polarity of `source` is assigned to the target, if ~ is specified in front of `source`.

- `target` is a sentence component `(SP| OP|PP) to which` the sentiment is directed.

After automatic detection of sentiment phrases in a sentence the system searches for matching the predicate structure in the sentiment pattern database and tries to identify the topic-sentiment.

Leveraging existing topic-based technologies is the most common practice for sentiment summarization. One line of practice is to adapt existing topic-based multi-document summarization algorithms to the sentiment setting. Sometimes the adaptation consists of simply modifying the input to these pre-existing algorithms. For instance, (Seki et. al., 2004) have proposed that one can apply standard multi-document summarization to a sub-collection of documents that are on the same topic and belong to some relevant genre of text, such as "*argumentative*".

(Pang and Lee, 2004) have proposed a two-step procedure for polarity classification for movie reviews, wherein they first detect the objective portions of a document (e.g., plot descriptions) and then apply polarity classification to the remainder of the document after the removal of these presumably uninformative portions. Importantly, instead of making the subjective-objective decision for each sentence individually, they postulate that there might be a certain degree of continuity in subjectivity labels (an author usually does not switch too frequently between being subjective and being objective), and incorporate this intuition by assigning preferences for pairs of nearby sentences to receive similar labels. All the sentences in the document are then labeled as being either subjective or objective through a collective classification process, where this process employs a reformulation of the task as one of finding a *minimum cut* in the appropriate graph. Two key properties of this approach are: (1) it affords the finding of an exact solution to the underlying

optimization problem via an algorithm that is efficient both in theory and in practice, and (2) it makes it easy to integrate a wide variety of knowledge sources about individual preferences that items may have for one or the other class and about the pair-wise preferences that items may have for being placed in the same class regardless of the particular class. Figure 5.1 illustrates the overview of the graph based minimum cut methodology.

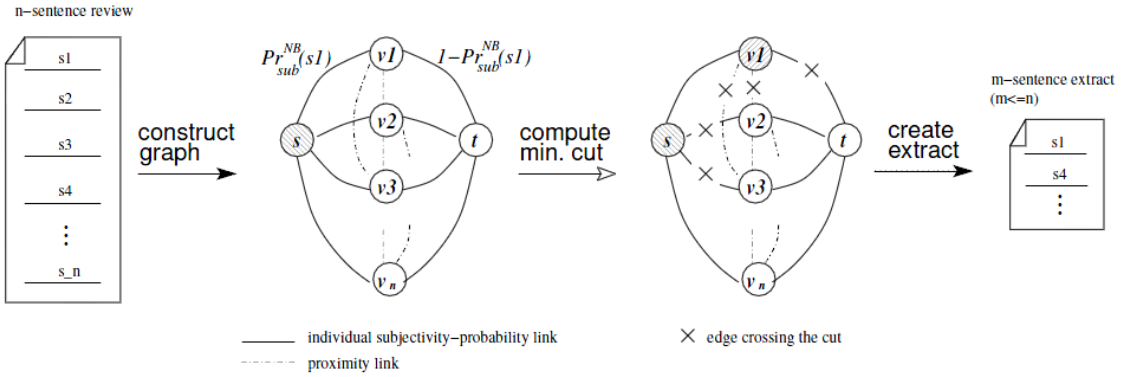


Figure 5.1: Graph-Cut-Based Creation of Subjective Extracts (Pang and Lee, 2004)

(Ku et. al., 2005) present techniques for automatic opinion summarization based on topic detection. The system selects representative words from a document set to identify the main concepts in the document set. A term is considered to represent a topic if it appears frequently across documents or in each document. The authors use many weighing mechanisms to detect the representative words (topic words) at sentence, paragraph or document level. The identified topics are then further used for opinion summarization.

(Zhou et. al., 2006) have proposed the architecture for generative summary from blogosphere. Typical multi-document summarization (MDS) systems focus on content selection followed by synthesis based on removing redundancy across multiple input documents. The online discussion summarization system works on an online discussion corpus that involves multiple participants and the discussion topics are passed back and forth by various participants. Due to the complex structure of the dialogue, similar subtopic structure identification in the participant-written dialogues is essential. Maximum Entropy Model (MEMM) and Support Vector Machine (SVM) have been used with a number of relevant features.

(Kawai et. al., 2007) developed a news portal site called Fair News Reader (FNR) that recommends news articles with different sentiments for a user in each of the topics in which the user is interested. FNR can detect various sentiments of news articles and determine the sentimental preferences of a user based on the sentiments of previously read articles by the user. News articles crawled from various news sites are stored in a database. The contents are integrated as needed and the summary is presented on one page. A sentiment vector on the basis of topic word lattice model has been generated for every document. The topic words are typically high frequent words of noun, adjective, adverb and verb category. The weighing mechanism plays a crucial role to identify the topic words from those high frequent words. A user sentiment model has been proposed based on the user sentiment state. The user sentiment state model works on the browsing history of the user.

The intersection of the documents under User Vector and Sentiment Vector are the recommended news articles for a particular user.

The Topic-Sentiment Mixture model proposed by (Mei et. al., 2007) is very effective for any real time output generation. A possible application of Topic-Sentiment analysis is shown in Figure 5.2. Given a query word that represent a user's ad-hoc information need (e.g., a product), the system extracts the subtopics in the search results, and associates each subtopic with positive and negative sentiments. From the example sentences on the left, which are organized in a two dimensional structure, the user can understand the *pros* and *cons* of each facet of the product, or what are its best and worst aspects.

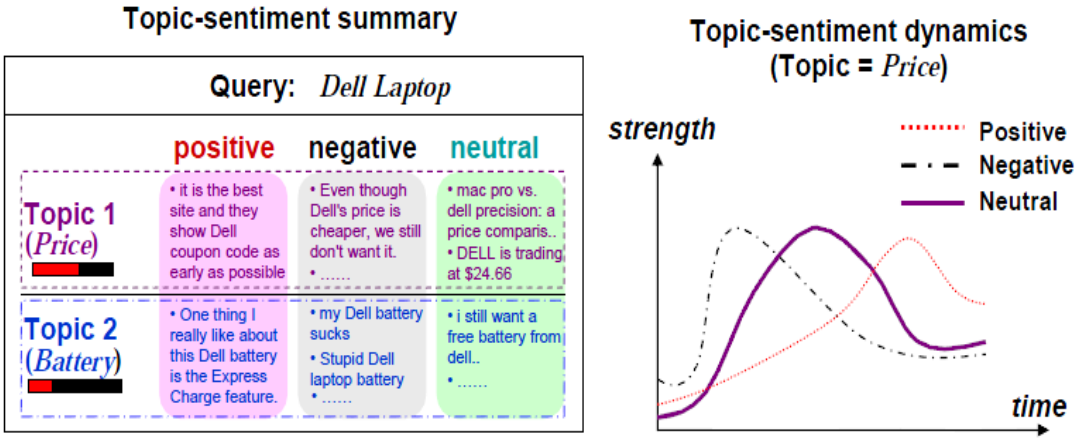


Figure 5.2: A Possible Application of Topic-Sentiment Analysis (Mei et. al., 2007)

5.1.2 Polarity-Wise

The topic-opinion model is the most popular one but there may be a requirement at the end users' perspective that they might look into an at-a-glance presentation of opinion-oriented summaries. For example: One market surveyor from company *A* might be interested in the root cause for why their product *X* (suppose camera) is becoming less popular day by day. *A* may want look into the *negative* reviews only. Relatively few research efforts could be found on the polarity-wise summarization in the literature compared to the popular topic-opinion model. Four important related works have been presented here which are significant in the aspects, problem definition and the solution architecture.

(Hu, 2004) has developed a review mining and summarization system that works in three steps: (1) mining product features that have been commented on by customers; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative and (3) summarizing the results. The proposed summary looks like the following in Figure 5.3. The positive and negative scores associated with each feature indicate the relative sentiment strength of the corresponding feature.

Digital_camera:

Feature: **picture quality**

Positive: 253

<individual review sentences>

Negative: 6

<individual review sentences>

Feature: **size**

Positive: 134

<individual review sentences>

Negative: 10

<individual review sentences>

…

Figure 5.3: An Example Summary Model Proposed by (Hu, 2004)

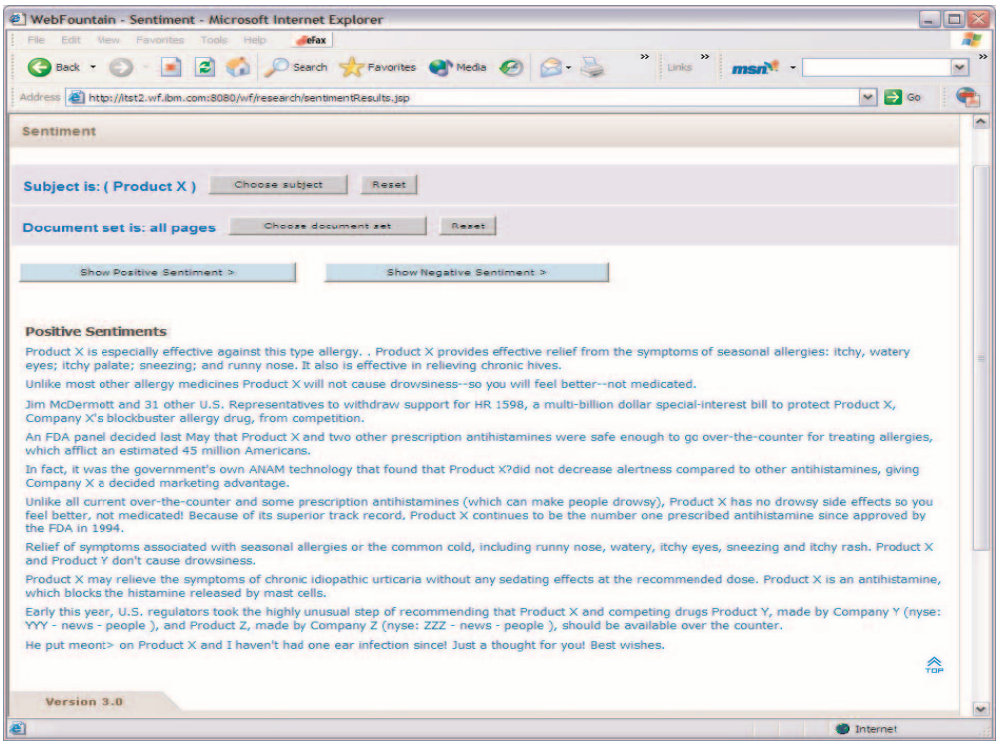


Figure 5.4: WebFountain System by (Yi and Niblack, 2005)

The sentiment mining system based on WebFountain has been proposed by (Yi and Niblack, 2005). The system extracts the opinionated sentences from a review corpus and presents them in an aggregated form based on their orientations such as positive or negative. The developed summary is presented in a GUI system as shown in Figure 5.4. Further details of the work are reported in the next section 5.1.3 on Visualization.

A multi-knowledge based approach for review mining and summarization has been proposed by (Zhuang et. al., 2006) which integrates WordNet, statistical analysis techniques and movie knowledge. All the sentences that express opinions on a movie feature class (e.g., cast, direction, star, category etc.) are collected. Then, the semantic orientation of the relevant opinion in each sentence is identified. Finally, the organized sentence list is shown as the summary. The following is an example of the polarity class wise summary produced by the system.

PRO (Positive): 70

Sentence 1: The movie is excellent.

Sentence 2: This is the best film I have ever seen.

…

CON (Negative): 10

Sentence 1: I think the film is very boring.

Sentence 2: There is nothing good with the movie.

…

(Das and Chen, 2007) have developed a methodology for extracting small investor sentiment from stock message boards. A total of five machine learning based classifiers have been used for opinion polarity classification and the final output is based on a voting scheme on the output produced by the different classifier algorithms. The five classifier algorithms are based on different approaches to message interpretation. Some of them are language independent while some are not.

Naive Classifier: This algorithm is based on a word count of positive and negative connotation words.

Vector Distance Classifier: If there are *D* words in the lexicon-dictionary, and each word is assigned a dimension in vector space, then the lexicon represents a *D-dimensional* unit hypercube. Every message may be thought of as a word vector $m \in R^D$ in this space and is therefore represented by a sparse vector.

Discriminant-Based Classifier: This is an updated version of the Naïve classifier based on positional importance.

Adjective-Adverb Phrase Classifier: This classifier is based on the assumption that adjectives and adverbs emphasize sentiment and require greater weight in the classification process. This algorithm

uses a word count, but restricts itself to words in specially chosen phrases containing adjectives and adverbs.

Bayesian Classifier: The classifier comprises of three components: (i) lexical words, (ii) message text, and (iii) classes or categories (bullish, bearish or neutral), resulting in the corpus standard word-message-class model. The Bayesian classifier uses word-based probabilities, and is thus indifferent to the structure of the language.

Final classification is based on achieving a simple majority vote amongst the five classifiers, i.e., three of five classifiers should agree on the message type. To produce the textual summary, four different metrics of classification performance, namely, percentage classification accuracy, percentage of false positives, percentage error in aggregate sentiment and a test of no classification ability have been considered. Finally, the summary is generated by voting for each polarity classes such as bullish, bearish or neutral.

5.1.3 Visualization

The graphical or the visualized output format is one of the trusted and well acceptable methods to convey all the automatically extracted knowledge to the end user in a concise format. A number of researchers have tried to leverage the existing or newly developed graphical visualization methods for the opinion summary presentation. Some important related works on opinion summary visualization techniques are now described.

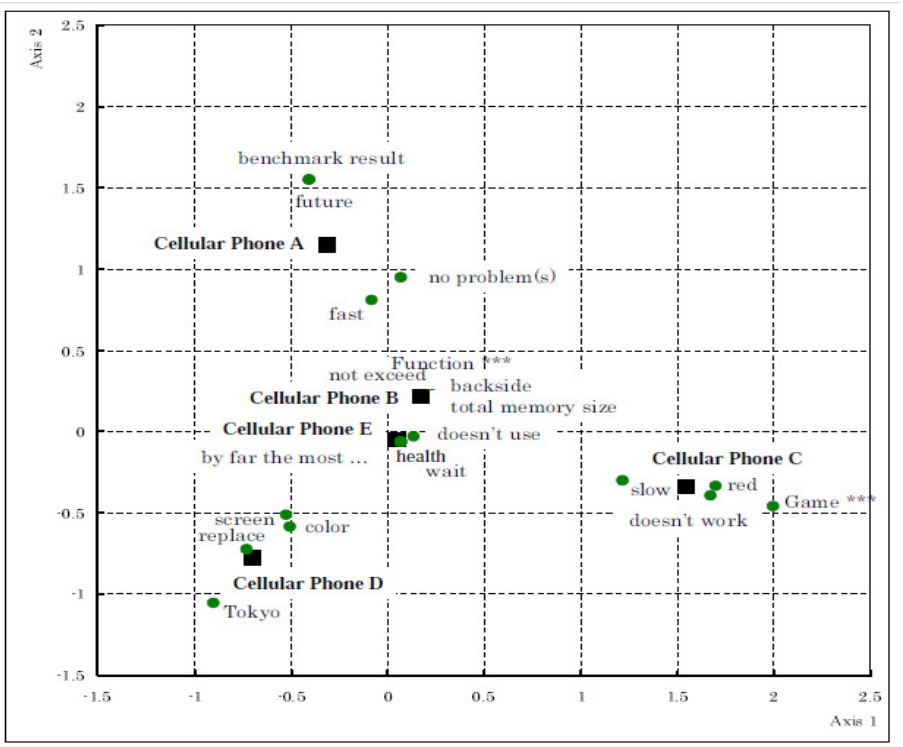


Figure 5.5: Positioning Map for Five Cellular Phones and their Extracted Characteristics by (Morinaga et al., 2002)

(Morinaga et. al., 2002) represent degrees of association between products and opinion-indicative terms of a polarity. First, opinions are collected using the system mentioned in (Tateishi et. al., 2001). Coding-length and probabilistic criteria are used to determine the terms to focus on, and principal component analysis is then applied to produce a two-dimensional visualization, such that nearness corresponds to strength of association (Li and Yamanishi, 2001). Thus, in Figure 5.5, cellphone *A* is associated with positive terms, whereas cellphone *C* is associated with negative terms.

(Gamon et. al., 2005) present a system called Pulse that extracts topic-sentiment information from customer written reviews, generally in free text format. The Pulse system displays the extracted information simultaneously into two dimensions, i.e., topic and sentiment. It first extracts taxonomies of major categories and minor categories of a particular topic (e.g., *cars*) by simply querying the review database. The sentences are then extracted from the reviews of each make and model and processed according to the two dimensions of information: sentiment and topic. To train the sentiment classifier, a small random selection of sentences is labeled by hand as expressing *positive*, *negative or other* sentiment. This small labeled set of data is used with the entirety of the unlabeled data to bootstrap a classifier. A *k-means* soft clustering has been implemented with *tf-idf* weighting. Once the sentences for a model of any topic have been assigned to clusters and have received a sentiment score from the sentiment classifier, the visualization component displays the clusters and the keyword labels that were produced for the sentences associated with the topic. Figure 5.6 shows the visualization from the Pulse system for the topic "*car*".

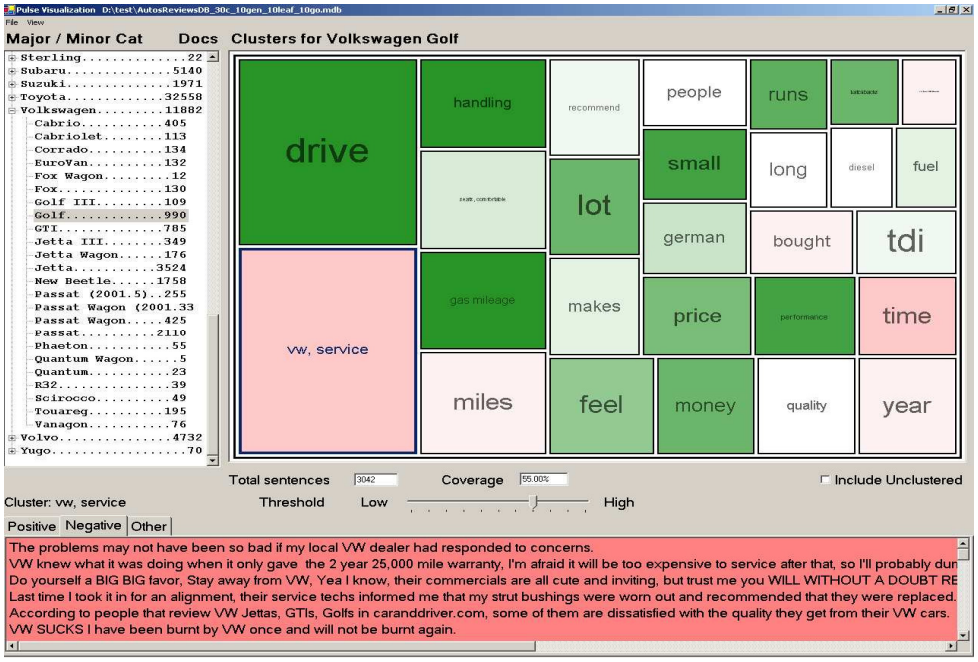


Figure 5.6: Screenshot of the Pulse user interface showing the taxonomy and the Tree Map with labeled clusters and sentiment coloring, and individual sentences from one cluster (for **Car**) (Gamon et. al., 2005)

(Yi and Niblack, 2005) have proposed several methodologies for sentiment extraction and visualization using WebFountain (Gruhl et. al., 2004). A *Sentiment Miner* system has been developed

with the basic backbone architecture of the WebFountain. The *Sentiment Miner* system is trained for both application-specific structured and unstructured data. A topic-sentiment model has been developed with the following hypothesis.

- A part-of relationship with the given topic.
- An attribute-of relationship (i.e., sub-topics) with the given topic.
- An attribute-of relationship with a known feature of the given topic.

The overall experiment has been done on product review dataset. The following Figure 5.7 shows the GUI output of the system in the pharmaceutical domain.

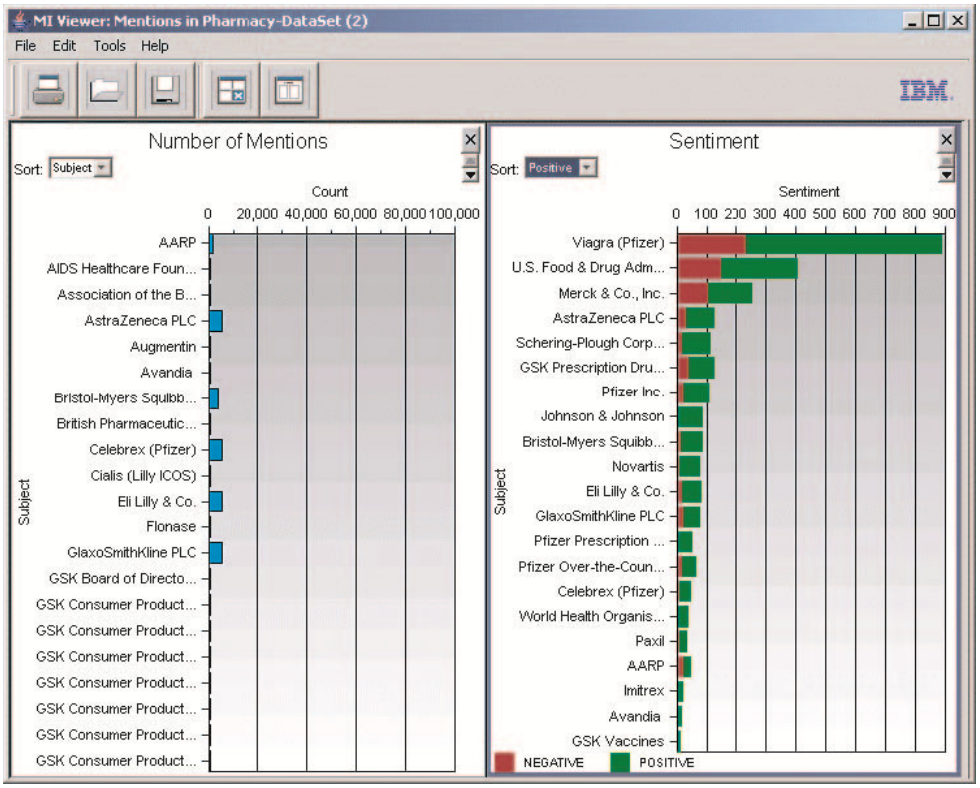


Figure 5.7: WebFountain: the GUI Visualization of the Sentiment Mining Result (Yi and Niblack, 2005)

(Carenini et. al., 2006)[3] present and compare two approaches to the task of multi document opinion summarization on evaluative texts. The first is a sentence extraction based approach while the second one is a natural language generation-based approach. Relevant extracted features are categorized into two types: User Defined Features (UDF) and Crude Features (CF) as described in (Hu and Liu, 2004). The authors also present a technique to present the summary in a multimedia visualization. It was hypothesized that the visualization component of the interactive multimedia summary should:

[3] http://www.cs.ubc.ca/~carenini/storage/SEA/demo.html

- Convey the user-defined hierarchical organization of the extracted knowledge.
- Communicate both the importance and the customer opinions about the extracted knowledge to the user.
- Allow the user to explore the original dataset.

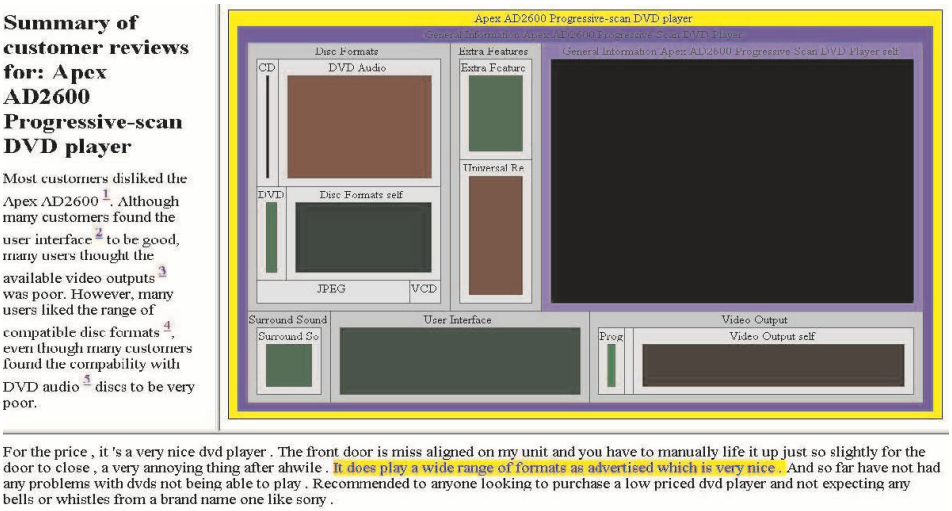


Figure 5.8: A Treemaps Visualization of Opinion Summary by (Carenini et. al., 2006)

The authors found that Treemaps (Shneiderman, 1992) could be adapted to satisfy all the three criteria. A Treemap is a two-dimensional space-filling technique for visualizing hierarchies. A Treemap represents an individual node in a tree as a rectangle with nested rectangles representing the descendants of the node. Because Treemaps use rectangles to represent trees, they can simultaneously visualize the hierarchy (the first criterion) and rapidly communicate other domain-specific information about each node by varying the size and fill color of the rectangles. These two dimensions can be naturally mapped into the domain: size can be used to represent the importance of a feature in the UDF while color can be used to represent customer opinions about a feature. This successfully fulfills the second criterion listed above. Figure 5.8 shows a screenshot of the interface to the Treemaps interactive summarizer. Each evaluation in the summary corresponds to a node in the Treemap, for example, "*available video outputs*" refers to the (non-leaf) node in the lower right corner. In the upper left part of the screen, the user sees the textual summary. A Treemap visualization occupies the majority of the upper part of the screen. The bottom of the screen provides space for the user to interactively access the text of the original reviews. In this image, the user has clicked on footnote 4, pointing her to a review in which the range of compatible discs is positively evaluated. The text of the review is shown in the bottom of the screen and the relevant sentence is highlighted.

The research efforts by (Gregory et. al., 2006) present techniques to extract and visualize the affective content of documents along with an interactive capability for exploring emotions in a large document collection. The proposed system first automatically identifies affective text by comparing each document against a lexicon of affect-bearing words and obtains an affect score for each

document. A number of visual *metaphors* have been proposed to represent the affect in the collection. A number of tools have been developed to interactively explore the affective content of the data.

In selecting *metaphors* to represent the affect scores of documents, the authors (Gregory et. al., 2006) started by identifying the kinds of questions that users would want to explore. A set of customer reviews for several commercial products were considered as the guiding example in (Hu and Liu, 2004). A user reviewing this data might be interested in a number of questions, such as:

- What is the range of overall affect?
- Which products are viewed most positively? Most negatively?
- What is the range of affect for a particular product?
- How does the affect in the reviews deviate from the norm? Which are more negative or positive than would be expected from the averages?
- How does the feedback of one product compare to that of another?
- Can we isolate the affect as it pertains to different features of the products?

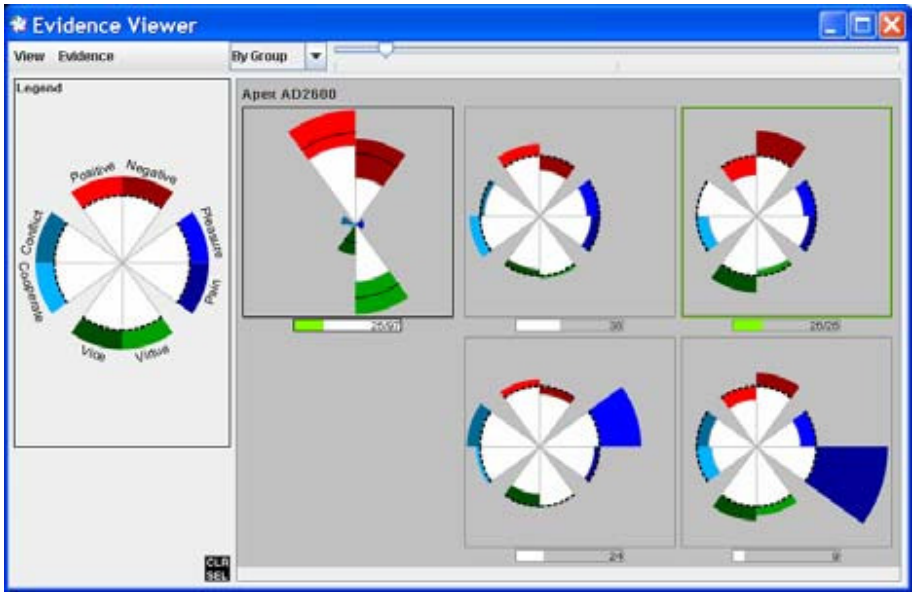


Figure 5.9: The Visualization of IN-SPIRE on Affect Summary (Gregory et. al., 2006)

(Gregory et. al., 2006) have used the IN-SPIRE (Hetzler and Turner, 2004) system which is a visual analytics tool designed to facilitate rapid understanding of large textual corpora. IN-SPIRE generates a compiled document set from mathematical signatures for each document in a set. Information is presented to the user using several visual *metaphors* to expose different facets of the textual data. The central visual metaphor is a *Galaxy* view of the corpus that allows users to intuitively interact with thousands of documents, examining them by theme. IN-SPIRE leverages the use of context

vectors such as Latent Semantic Analysis (LSA) (Deerwester et. al., 1990) for document clustering and projection. Additional analytic tools allow exploration of temporal trends, thematic distribution by source or other metadata, and query relationships and overlaps. IN-SPIRE has been enhanced to support visual analysis of sentiment as shown in the Figure 5.9. For the current visualization, the authors drew inspiration from the Rose plot used by Florence Nightingale (Wainer, 1997).

5.1.4 Tracking

In many applications, analysts and other users are interested in tracking changes in sentiment about a product, political candidate, company or other issues over time. The tracking system could be a good measure to understand the people's sentiment changes, e.g., in sociological surveys. In general sense, tracking means plotting of sentiment values over time into a graphical visualization.

The Lydia[4] project (also called TextMap) (Lloyd et. al., 2005) seeks to build a relational model of people, places and many more other things through natural language processing of news sources and the statistical analysis of entity frequencies and co-locations. The system tracks the temporal and spatial distribution of the entities in the news: *who* is being talked about, by *whom*, *when* and *where*? The Lydia system relies on visual output and the previously mentioned aspects are reported by the *juxtapositional*, *spatial* and the *temporal* entity analysis.

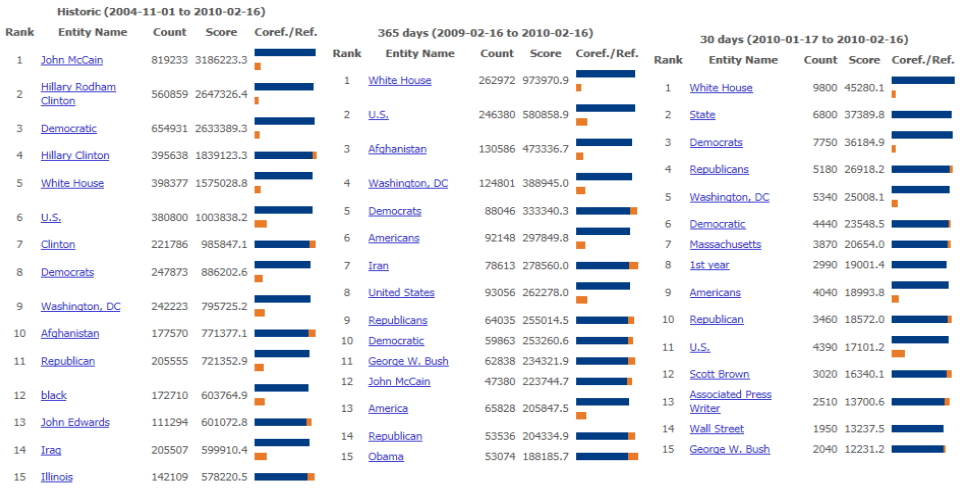


Figure 5.10: Juxtaposition Analysis by Lydia for "*Barack Obama*" (Lloyd et. al., 2005)

Juxtaposition: The entity about which or whom any news text is written fits into the world largely depending on how it relates to other entities. For each entity, the Lydia system computes a significance score for every other entity that co-occurs with it and ranks its juxtapositions by this score. A visual illustration of the juxtaposition analysis by Lydia for "*Barack Obama*" is reported in the Figure 5.10.

[4] http://www.textmap.com/

Spatial Analysis: Each newspaper has a location and a circulation and each city has a population. Based on these facts the Lydia approximates a *sphere of influence* for each newspaper and the particular desired topic and finally visualize the information with a color intensity marking in a geographic map to illustrate the people's sentiment over the geographic locations about the particular topic. A spatial analysis by Lydia on "*Barack Obama*" is reported in the Figure 5.11.

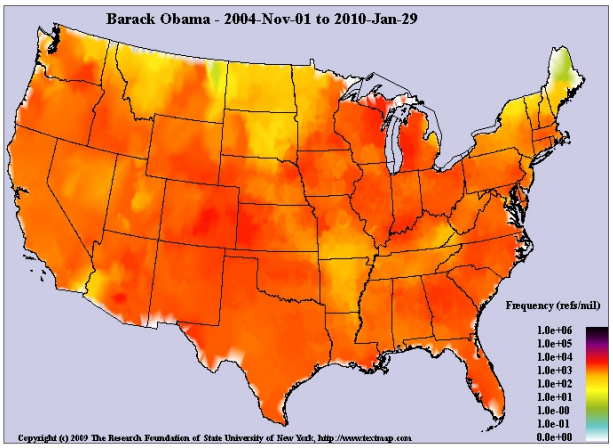


Figure 5.11: Where is "*Barack Obama*" HOT? (Lloyd et. al., 2005)

Temporal Analysis: Every published news item has a publication date. Based on the temporal information the people's sentiment has been plotted in a graphical output by Lydia. The temporal sentiment analysis on "*Barack Obama*" is reported in the Figure 5.12.

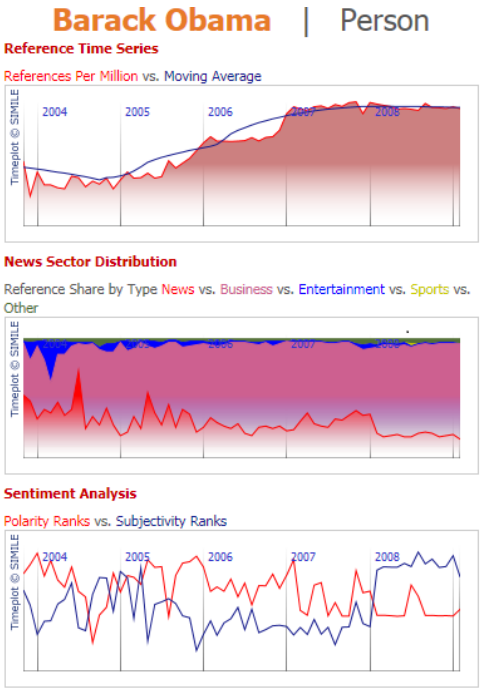


Figure 5.12: Temporal Analysis for "*Barack Obama*" by Lydia (Lloyd et. al., 2005)

(Ku et. al., 2006) hypothesize that ***opinion extraction***, ***opinion summarization*** and ***opinion tracking*** are three important techniques for understanding opinions. Opinion extraction mines opinions at word, sentence and document levels from articles. Opinion summarization summarizes opinions of articles by identifying the sentiment polarities, the degree and the correlated events. Opinion tracking visually reports the opinion changes over time. The authors investigated their techniques on both the news and web blog articles. TREC[5] and NTCIR[6] articles are collected from the web blogs and these articles serve as the information sources for this task. A detailed description of the corpus development and annotation process has also been reported. The visual representation of the opinions tracking system for four persons who participated in the Presidential election in Taiwan in March 2000 is shown in the Figure 5.13.

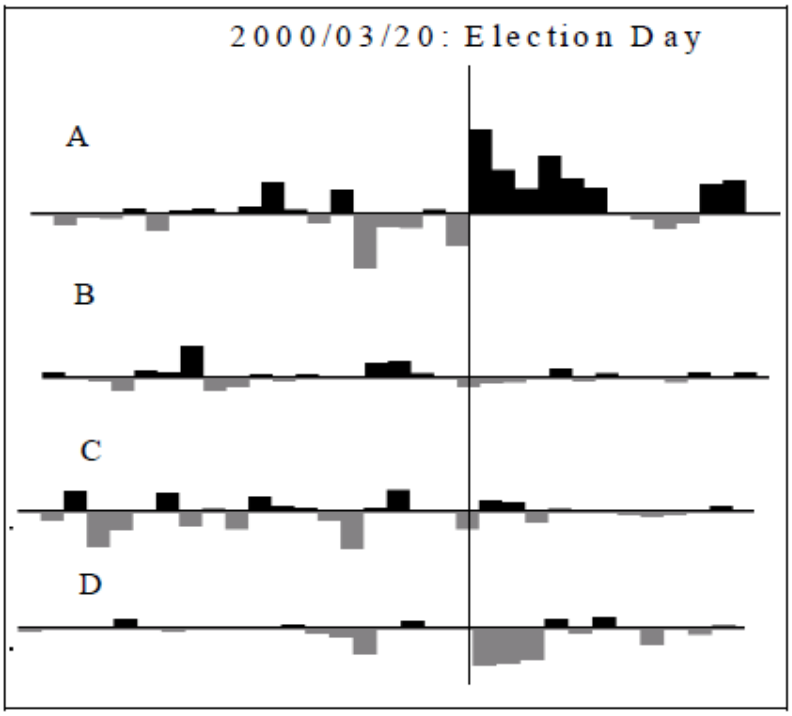


Figure 5.13: Opinions Tracking for Four Electoral Candidates (Ku et. al., 2006)

Persons *A*, *B* and *C* were candidates and *D* was the president at that time. The trend fits the opinions in this period and the opinion summaries can tell events correlated with these opinions. This tracking system can also track opinions according to different requests and different information sources, including news agencies and the web. Opinion trends toward one specific focus from different opinions expressed by various people can also be compared. This information is very useful for the government, institutes, companies and the concerned public.

[5] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG
[6] http://research.nii.ac.jp/ntcir/index-en.html

(Mishne and Rijke, 2006) demonstrate a system for tracking and analyzing the moods of bloggers worldwide. The demonstrated system is trained on the largest blogging community, LiveJournal[7]. Users of LiveJournal, currently the largest weblog community, have the option of reporting their mood at the time of the post; users can either select a mood from a predefined list of 132 common moods such as "amused" or "angry," or enter free-text. The authors developed a system, called MoodViews[8], a collection of tools for analyzing, tracking and visualizing moods and mood changes in blogs posted by LiveJournal users. MoodViews consists of three components, each offering a different view of global mood levels, the aggregate across all postings of the various moods: ***Moodgrapher*** tracks the global mood levels, ***Moodteller*** predicts them, and ***Moodsignals*** helps in understanding the underlying reasons for mood changes. A brief presentation of each of these services is now reported.

Moodgrapher: the basic component of the system plots the aggregate mood levels over time. Sample plots, showing irregular mood patterns following events plotted by *Moodgrapher*, are shown in the Figure 5.14.

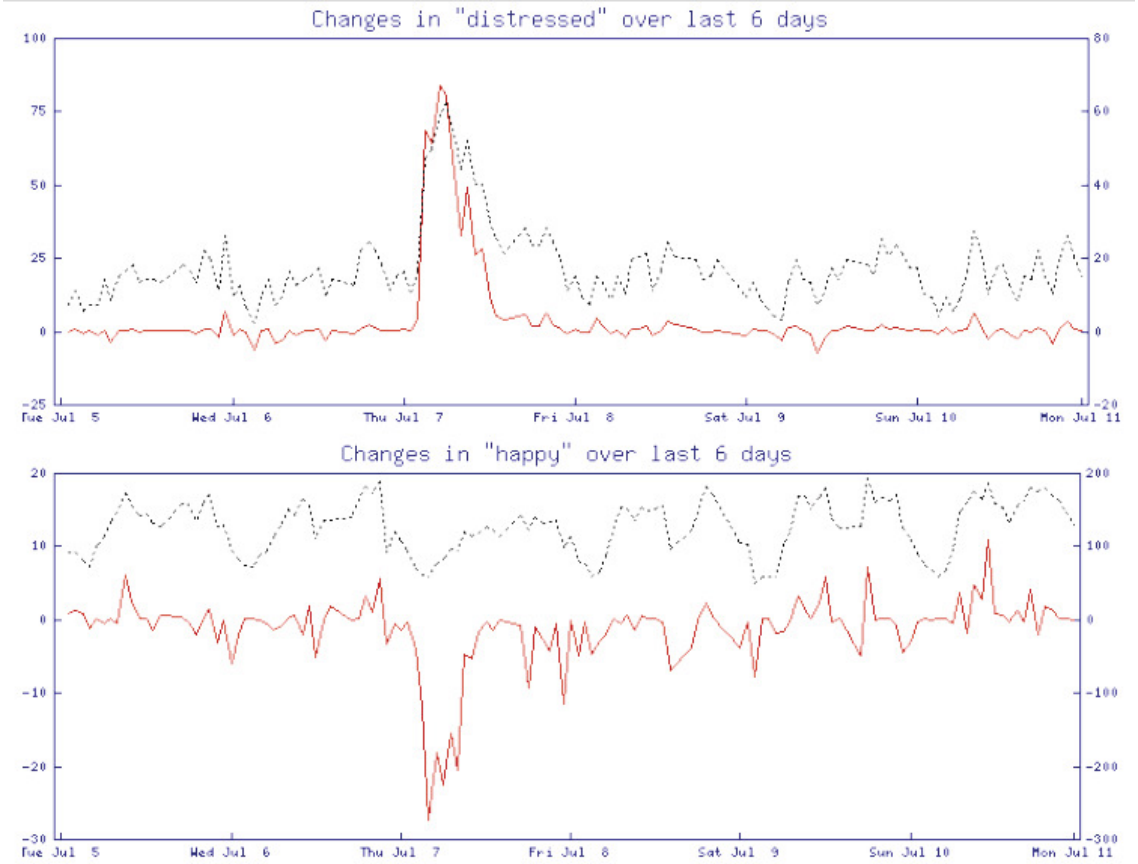


Figure 5.14: Global Moods Plotted by Moodgrapher: Distress Peaks and Happiness Plunges after Terrorists Strike London on July 7, 2005, (Mishne and Rijke, 2006)

[7] http://www.livejournal.com/
[8] http://moodviews.com/

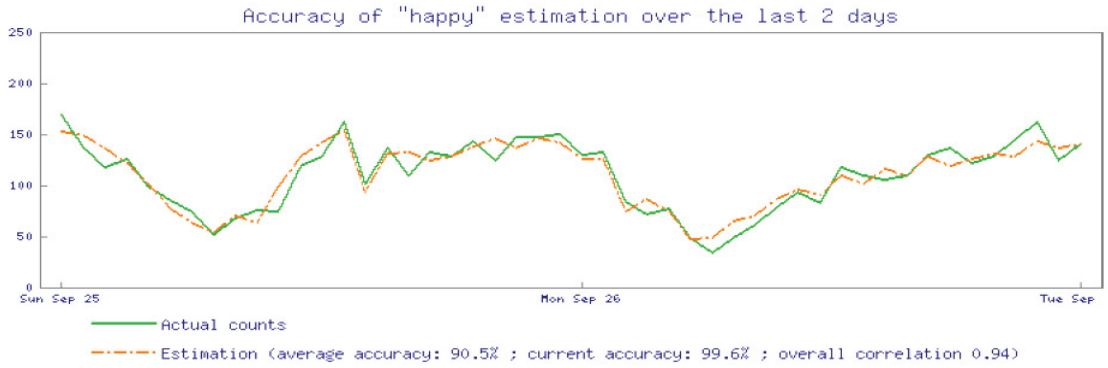


Figure 5.15: Moodteller in Action: Estimating "*happiness*" over Two Days at the end of September 2005 (Mishne and Rijke, 2006)

Moodteller: *Moodteller* goes a step beyond *Moodgrapher*, and uses natural language processing and machine learning techniques to estimate the mood levels from the text of blog entries posted on LiveJournal, without using the mood tags provided by bloggers. An example of *Moodteller* prediction plotted by the *Moodgrapher* is reported in the Figure 5.15.

Moodsignals: Users of *Moodgrapher* witnessing irregular behavior, such as a spike in a certain mood, are often interested in discovering the cause of this spike—typically, an event affecting a large number of people. *Moodsignals* detects words and phrases which are associated with a given mood in a given time interval. With the *Moodsignals*, users can simply select a region of a mood graph they are interested in, and view a ranked list of the terms most related to the mood at that time. An example is shown in the Figure 5.16.

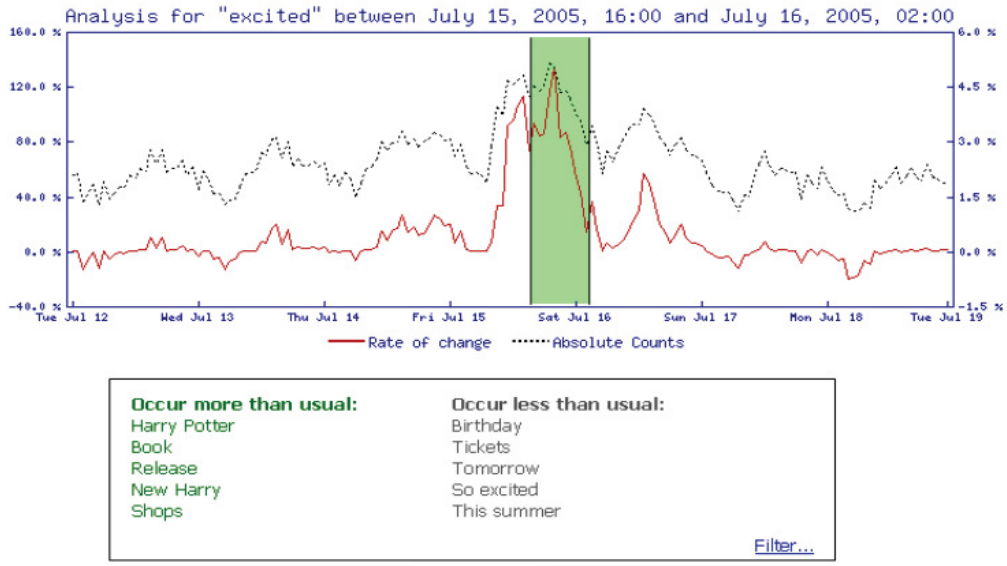


Figure 5.16: Moodsignals Uncovering the Excitement Peak on July 16, 2005: The Release of a new Harry Potter Book (Mishne and Rijke, 2006)

(Fukuhara et. al., 2007) present their research efforts on finding the temporal sentiment analysis that analyzes temporal trends of sentiments and topics from a text archive. The system accepts texts with timestamp such as Weblog and news articles, and produces two kinds of graphs, i.e., (1) ***topic graph*** that shows ***temporal change*** of topics associated with a sentiment, and (2) ***sentiment graph*** that shows ***temporal change*** of sentiments associated with a topic. Figure 5.17 illustrates the overview of the proposed problem definition called ***temporal sentiment analysis***.

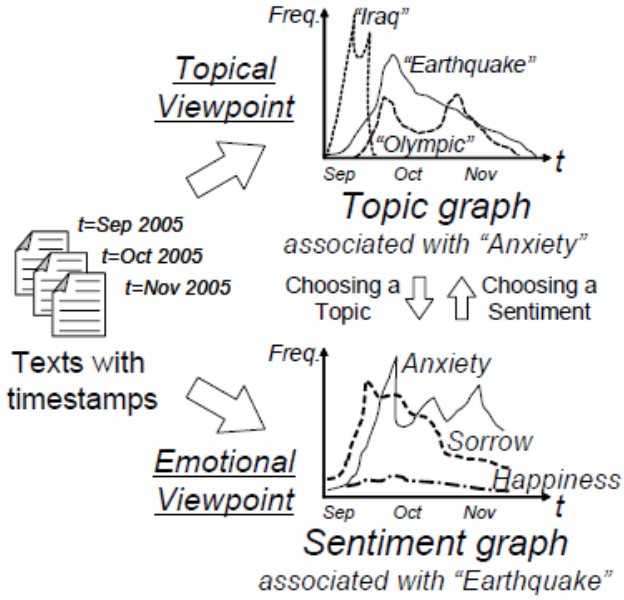


Figure 5.17: Temporal Sentiment Analysis (Fukuhara et. al., 2007)

The following is the procedure for making a topic graph.

Given: one sentiment s from the sentiment category *S* and the period of time: $D = (d_1, d_2, \ldots, d_l)$ specified by a user

Step 1: For each day d_i in *D*, retrieve articles containing sentiment phrases of sentiment *s*.

Step 2: Extract keywords from retrieved articles by using a keyword extraction system called GENSEN-Web[9] that can extract compound nouns as a keyword.

Step 3: For each extracted keywords w_j $(j=1,2,\ldots,N)$, calculate an average correlation *c* between w_j and sentiment phrases contained in *S*. The Dice coefficient has been used for calculating the correlation.

Step 4: Extract top *n* keywords according to the score defined by the products of (1) number of days in which keywords appears, (2) inverse frequency of number of days, and (3) scores provided by GENSEN-Web.

[9] http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html

Step 4'(optional): Put keywords into clusters based on correlation coefficient over timeline and the Dice coefficient in an article.

Step 5: Generate a temporal graph for each n keywords (or clusters). For viewability of the graph, the moving average has been calculated.

An example topic graph for sentiment tracking is shown in the Figure 5.18.

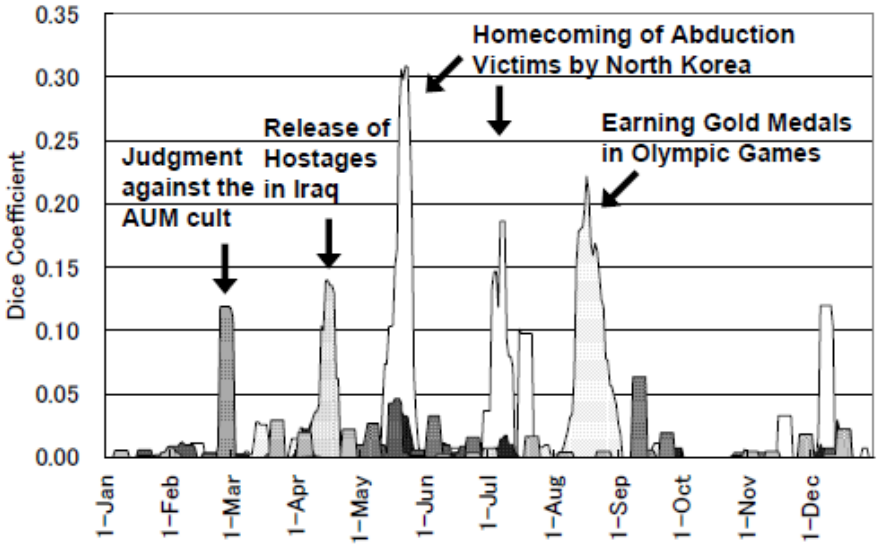


Figure 5.18: Topic Graph for Sentiment "*happy*" in 2004 (Using Clustering Option), (Fukuhara et. al., 2007)

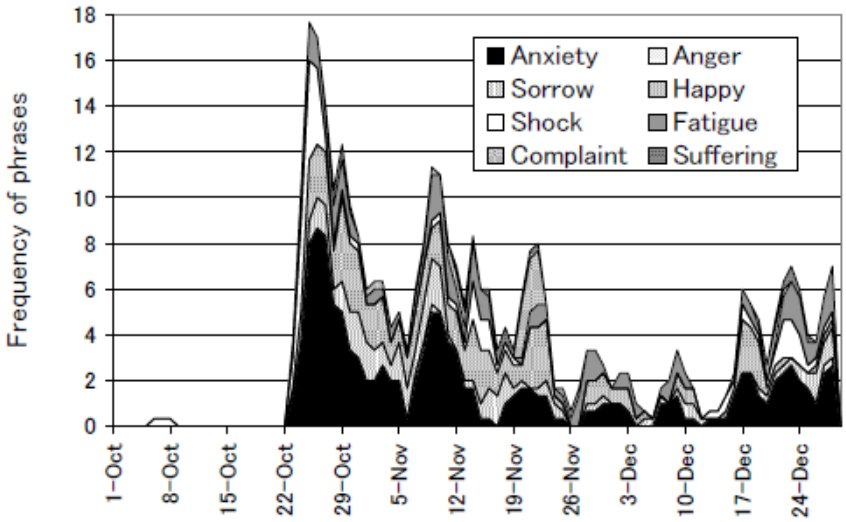


Figure 5.19: Sentiment Graph for the Topic "*earthquake*" in the Fourth Quarter in 2004 (Stacked Chart), (Fukuhara et. al., 2007)

The following is the procedure for making a sentiment graph.

Given: a keyword w specified by a user along with a period of time: $D = (d_1, d_2, \ldots, d_l)$

Step 1: Retrieve articles containing keyword *w* for each day d_i $(i=1, 2,\ldots, l)$.

Step 2: For each article, calculate the sum of frequencies of sentiment phrases for all sentiment categories.

Step 3: Generate a temporal graph on the frequency of sentiment phrases for each sentiment category. Then, moving average is applied to the graph.

An example sentiment graph has been reported in the Figure 5.19.

5.2 The Proposed 5W Sentiment Summarization-Visualization-Tracking System

It has been observed during the literature survey that no consensus among the researchers could be found on the output format of any opinion summarization system. Actually the output format varies on the end user's requirement and the domain for which the system has been tuned. In the present work, experiments have been carried out with multiple output formats. First the experiment started with the multi-document topic-opinion textual summary (Das and Bandyopadhyay, 2010(j));(Das and Bandyopadhyay, 2010(k)) but realizing the end user's requirements, *at-a-glance* presentation of accumulated information and the 5W constituent based textual summarization, visualization and tracking system has been devised (Das and Bandyopadhyay, 2012(d)). The 5W constituent based summarization system is a multi-genre system that supports both the most acceptable output generation: textual and visualization format. The system facilitates users to generate sentiment tracking with textual summary and sentiment polarity wise graph based on any dimension or combination of dimensions as they want, for example, "*Who*" are the actors and "*What*" are their sentiment regarding any topic, changes in sentiment during "*When*" and "*Where*" and the reasons for change in sentiment as "*Why*". The 5W constituent based summarization system falls into every genre, "*Topic-Wise*", "*Polarity-Wise*" or "*Other-Wise*".

Topic-Wise: The system facilitates users to generate sentiment summary based on any customized topic like Who, What, When, Where and Why based on any dimension or combination of dimensions they want.

Polarity-Wise: The system produces an overall gnat chart that can be treated as the overall polarity wise summary. An interested user can still look into the summary text to find out more details.

Visualization and Tracking: The visualization facilitates users to generate visual sentiment tracking with polarity wise graph based on any dimension or combination of dimensions as they want, i.e., "Who" are the actors and "What" are their sentiment regarding any topic, changes in sentiment

during "When" and "Where" and the reasons for change in sentiment as "Why". The final graph for tracking is generated with a timeline.

Moreover the end user can structure their information need as:

- ***Who***? Who was involved?
- ***What***? What happened?
- ***When***? When did it take place?
- ***Where***? Where did it take place?
- ***Why***? Why did it happen?

During the development of the multi-document topic-opinion summarization system, a strong semantic lexical network has been proposed, following the idea of Mental Lexicon models (Ferret and Zock, 2006). The same lexical semantic network has been used to develop the 5W system. The 5W structurization system, as discussed in the chapter 4 has been involved to extract the 5W semantic roles from the opinionated sentences. The development process of the Multi-Document Topic-Opinion summarizer has been described in the section 5.3 and the 5W sentiment Summarization-Visualization-Tracking system has been discussed in section 5.4. Actually the 5W sentiment Summarization-Visualization-Tracking uses the core components of the Multi-Document Topic-Opinion summarizer.

5.3 Multi-Document Topic-Opinion Extractive Summary

In this section the development of an opinion summarization system that works on Bengali News corpus has been described. The system identifies the sentiment information in each document, aggregates them and represents the summary information in text. The present system follows a topic-sentiment model for sentiment identification and aggregation. Topic-sentiment model is designed as discourse level *theme* identification and the topic-sentiment aggregation is achieved by *theme clustering (k-means)* and Document level Theme Relational Graph representation. The Document Level Theme Relational Graph is finally used for candidate summary sentence selection by standard *page rank* algorithms used in Information Retrieval (IR). As Bengali is a resource constrained language, the building of annotated gold standard corpus and acquisition of linguistics tools for lexico-syntactic, syntactic and discourse level features extraction are described in the following sub-sections.

5.3.1 Corpus

For the present task a Bengali news corpus has been developed from the archive of a leading Bengali news paper available on the Web (`http://www.anandabazar.com/`). A portion of the corpus from the editorial pages, i.e., Reader's opinion section or Letters to the Editor Section containing 28K

word forms has been manually annotated with sentence level subjectivity and discourse level theme words.

5.3.2 Annotation

From the collected document set (Letters to the Editor Section), some documents (containing 28K word forms) have been chosen for the annotation task. Some statistics about the Bengali news corpus is represented in the Table 5.1. Documents that have appeared within an interval of four months are chosen on the hypothesis that these letters to the editors will be on related events. A simple annotation tool (a snapshot is shown in Figure 5.20) has been designed for annotating the subjective sentences. The tool highlights the sentiment words (based on the occurrence of the word in the SentiWordNet (Bengali), described in the Chapter One) by two different colors within a document according to their sentiment orientation categories (GREEN: Positive words, RED: Negative words as reported in the Figure 5.20). The tool also highlights the title words (YELLOW) and theme words (BLUE), automatically identified by the rule-based theme detection technique (described in section 2.6.1). For example the words "নরেন্দ্র মোদি" and "ঘুষ" are the title words, i.e. occurs in title of the document thus highlighted in yellow. Words like "গুজরাট" and "মামলার" are the theme words thus highlighted in blue. The words highlighted in either green ("অভিযোগ" and "প্রভাবিত") or red ("স্বীকার") are the sentiment words extracted from SentiWordNet (Bengali) (Das and Bandyopadhyay, 2010(f))[10] .

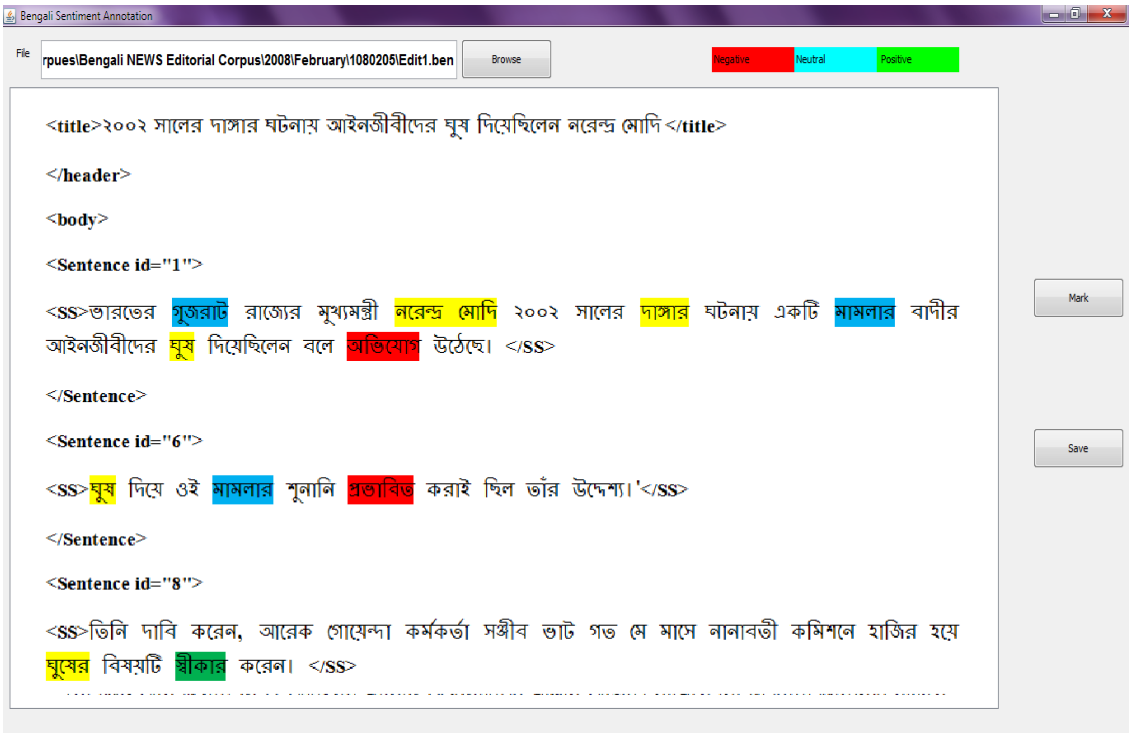


Figure 5.20: The Subjectivity Annotation Tool for Bengali

[10] http://www.amitavadas.com/sentiwordnet.php

The documents with such annotated sentences are saved in XML format. The XML tag "<SS>" stands for subjective sentence as shown in fig 5.21.

```
<document docid="Modi-324" encoding="UTF-8">

<header>

<title>২০০২ সালের দাঙ্গার ঘটনায় আইনজীবীদের ঘুষ দিয়েছিলেন নরেন্দ্র মোদি </title>

</header>

<body>

<Sentence id="1">

<SS>ভারতের গুজরাট রাজ্যের মুখ্যমন্ত্রী নরেন্দ্র মোদি ২০০২ সালের দাঙ্গার ঘটনায় একটি মামলার বাদীর আইনজীবীদের ঘুষ দিয়েছিলেন বলে অভিযোগ উঠেছে। </SS>

</Sentence>

<Sentence id="6">

<SS>ঘুষ দিয়ে ওই মামলার শুনানি প্রভাবিত করাই ছিল তাঁর উদ্দেশ্য।'</SS>

</Sentence>

<Sentence id="7">

দাঙ্গার ঘটনা তদন্তে গঠিত নানাবতী কমিশনে দেয়া গোয়েন্দা কর্মকর্তা শ্রীকুমারের নথির একটি কপিও সাংবাদিকদের কাছে সরবরাহ করেন মল্লিকা।

</Sentence>

<Sentence id="8">

<SS>তিনি দাবি করেন, আরেক গোয়েন্দা কর্মকর্তা সঞ্জীব ভাট গত মে মাসে নানাবতী কমিশনে হাজির হয়ে ঘুষের বিষয়টি স্বীকার করেন। </SS>

</Sentence>
```

Figure 5.21: Subjectivity Annotation XML Format for Bengali

Bengali NEWS Corpus Statistics	
Total number of documents in the corpus	100
Total number of sentences in the corpus	2234
Average number of sentences in a document	22
Total number of wordforms in the corpus	28807
Average number of wordforms in a document	288
Total number of distinct wordforms in the corpus	17176

Table 5.1: Statistics of Bengali Sentiment Summarization-Tracking Corpus

Annotators were asked to annotate sentences for summary and to mark the theme words (topical expressions) in those sentences. The documents with such annotated sentences are saved in XML format. Figure 5.21 shows the XML annotation format. "*<SS>*" marker denotes subjective sentences and "*<TW>*" denotes the theme words. An English gloss has been added in the Figure 5.21 for readability but in the actual case only the Bengali sentences are marked.

5.3.3 Inter-Annotator Agreement

The agreement of annotations among the three annotators has been evaluated. Three annotators (Mr. X, Mr. Y and Mr. Z) participated in the present task. The agreement of tag values at theme words level and sentence levels are listed in Tables 5.2 and 5.3 respectively.

Annotators	X vs. Y	X Vs. Z	Y Vs. Z	Avg.
Percentage	82.64%	71.78%	80.47%	78.30%
All Agree	69.06%			

Table 5.2: Inter-Annotator Agreement at Theme Words Level

Annotators	X vs. Y	X Vs. Z	Y Vs. Z	Avg.
Percentage	73.87%	69.06%	60.44%	67.8%
All Agree	58.66%			

Table 5.3: Inter-Annotator Agreement at Subjective Sentence Level

From the analysis of inter-annotator agreement statistics, it is observed that the agreement drops fast as the number of annotators increases. It is less possible to have consistent annotations when more annotators are involved. In the present task, the inter-annotator agreement is better for theme words annotation rather than candidate sentence identification for summary. A small number of documents have been considered.

Further discussion with annotators reveals that the psychology of annotators is to identify as many possible theme words during annotation but the same groups of annotators are more cautious during sentence identification for summary as they are very conscious to find out the most concise set of sentences that best describe the opinionated snapshot of any document. The annotators were working independent of each other and they were not trained linguists.

5.3.4 Theme Detection

Term Frequency (TF) plays a crucial role to identify document relevance in Topic-Based Information Retrieval. The motivation behind developing the Theme detection technique is that in many documents relevant words may not occur frequently or irrelevant words may occur frequently. Moreover for sentiment analysis, theme words should have sentiment orientation. The Theme detection technique has been proposed to resolve these issues to identify discourse level relevant topic-semantic nodes in terms of word or expressions using a standard machine learning technique.

The machine learning technique used here is Conditional Random Field (CRF)[11]. The theme word detection is defined as a sequence labeling problem. Depending upon the input features, each word is tagged as either Theme Word (TW) or Other (O).

5.3.5 Feature Organization for Theme Detection

Features are basically the linguistic clues to detect the desired pattern of themes and the clues may exist at any level - lexical, syntactic or discourse. Feature engineering involves identification of best features followed by feature identification and feature extraction. It plays a crucial role in any kind of NLP task. In the theme detection task, the aim is to find out the concise and effective set of features. Identification and extraction of more semantically rich features for the Bengali language demand sophisticated linguistic tools, which is still unavailable. To overcome this problem, the best feature list is identified using the available tools or the features extracted by developing least complex modules. Theme refers to the sentimental topic of a document; therefore the system identifies themes as a bag-of-words involving both the linguistics and sentiment clues to identify it. The complete list of lexical, syntactic and discourse level feature sets are reported in the Table 5.4.

Level	Features
Lexical	POS
	SentiWordNet
	Frequency
	Stemming
Syntactic	Chunk Label
	Dependency Parsing Depth
Discourse	Title of the Document
	First Paragraph
	Term Distribution
	Collocation

Table 5.4: Features for Theme Detection

Once the best feature set has been identified then the next challenge is to extract those features effectively. Various linguistics tools that are used to extract the features are reported in the following sub-sections.

5.3.5.1 Lexical Features

Topic-opinion identification involves semantic understanding. Lexical features are the basic linguistic clues to identify the semantic role of any predicate. The following features have been experimentally identified as the effective features for the present task.

5.3.5.1.1 Part of Speech (POS)

It has been shown in (Hatzivassiloglou et. al., 2000; Chesley et. al., 2006) etc. that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion-topic identification

[11] http://crfpp.sourceforge.net

systems, like (Nasukawa et. al., 2003) are based on adjective or adverb words. The tool that has been used here is the Bengali Shallow Parser[12] developed under Indian Languages to Indian Languages Machine Translation (IL-ILMT) project funded by Department of Information Technology, Government of India.

5.3.5.1.2 SentiWordNet (Bengali)

Words that are present in the SentiWordNet carry opinion information. The presence of a Bengali word in the developed SentiWordNet (Bengali) (Das and Bandyopadhyay, 2010(f)) is used as an important feature during the learning process. These features associate the individual sentiment words or word n-grams (multiword entities) with strength measure as strong subjective or weak subjective. Strong and weak subjective measures are treated as a binary feature in the supervised classifier. Words which are collected directly from SentiWordNet (Bengali) are tagged with positivity or negativity scores. The subjectivity score of these words are calculated as:

$$E_s = | S_p | + | S_n |$$

where E_s is the resultant subjective measure and S_p, S_n are the positivity and negativity scores respectively.

5.3.5.1.3 Frequency

Frequency always plays a crucial role in identifying the importance of a word in the document. The system generates four separate high frequent word lists for four POS categories: adjective, adverb, verb and noun after function words are removed. Word frequency values are then effectively used as a crucial feature in the Theme Detection technique.

5.3.5.1.4 Stemming

Several words in a sentence that carry topic-opinion information may be present in inflected forms and stemming is necessary for them before they can be searched in the appropriate lists. Due to non availability of good stemmers in Indian languages especially in Bengali, a stemmer (Das and Bandyopadhyay, 2010(l)) based on stemming cluster technique has been used. This stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center (discussed in the Appendix).

5.3.5.2 *Syntactic Features*

Syntactic features depict the topic-opinion behavior of any chunk / phrase. The following syntactic features are used in the present system.

[12] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

5.3.5.2.1 Chunk Label

Chunk level information is effectively used as a feature in the supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. The tool that has been used here is Bengali Shallow Parser developed under Indian Languages to Indian Languages machine Translation (IL-ILMT) project funded by Department of Information Technology, Government of India.

5.3.5.2.2 Dependency Parser

Dependency depth feature is very useful to identify Theme expressions. A particular Theme word generally occurs within a particular range of depths (e.g., level 3-5 in a dependency tree) in a dependency tree. Theme expressions may be a Named Entity (NE: person, organization or location names), a common noun (e.g., accident, bomb blast, strike etc) or words of other POS categories. It has been observed that depending upon the nature of Theme expressions it can occur within a certain depth in the dependency tree for the sentence. A statistical dependency parser has been used for Bengali as described in (Ghosh et. al., 2009).

5.3.5.3 *Discourse Level Features*

It has been shown by various researchers (Somasundaran, 2010; Polanyi et. al., 2004) that sentimental semantics heavily depends on discourse level relations. No tool for identifying discourse level relations is publicly available for Bengali. Very simple and generic discourse level features have been used in the present work. As these features are very simple and generic in nature, they could be easily identified for any new language.

5.3.5.3.1 Positional Aspect

Depending upon the position of subjectivity clue, every document is divided into a number of zones. Various values of this feature are Title of the document, the first paragraph and the last two sentences. A detailed study was carried out on the Bengali corpus to identify the roles of the positional aspect (first paragraph, last two sentences) in the sentence level subjectivity detection task. It has been observed that generally first paragraph and last two sentences of any document contain subjectivity. Corpus statistics prove the phenomenon as reported in the Table 5.5. In 56.8% cases of the first paragraph in Bengali corpus respectively carry subjective information, whereas in 78.0% cases of last two sentences in Bengali corpus respectively carry subjective information.

Positional Factors	Bengali
First Paragraph	56.80%
Last Two Sentences	78.00%

Table 5.5: A Corpus Statistics on Document Level Positional Aspect of the Subjective Sentences

5.3.5.3.1.1 Title Words

Title words of a document always carry some meaningful thematic information. The title word feature has been used as a binary feature during CRF based machine learning.

5.3.5.3.1.2 First Paragraph Words

People usually give a brief idea of their beliefs and speculations in the first paragraph of the document and subsequently elaborate or support them with relevant reasoning or factual information. Hence, first paragraph words are informative in the detection of Thematic Expressions.

5.3.5.3.1.3 Words from Last Two Sentences

Generally every document concludes with a summary of the opinions expressed in the document.

5.3.5.3.2 *Term Distributional Model*

An alternative to the classical TF-IDF weighting mechanism of standard Information Retrieval (IR) has been proposed as a model for the distribution of a word. The model characterizes and captures the informativeness of a word by measuring how regularly the word is distributed in a document. (Carenini et. al., 2006) have introduced the opinion distribution function feature to capture the overall opinion distributed in the corpus. Thus the objective is to estimate $f_d(w_i)$ that measures the distribution pattern of the k occurrences of the word w_i in a document *d*. Zipf's law describes distribution patterns of words in an entire corpus. In contrast, term distribution models capture regularities of word occurrence in subunits of a corpus (e.g., documents, paragraphs or chapters of a book). A good understanding of the distribution patterns is useful to assess the likelihood of occurrences of a word in some specific positions (e.g., first paragraph or last two sentences) of a unit of text. Most term distribution models try to characterize the informativeness of a word identified by inverse document frequency (IDF). In the present work, the distribution pattern of a word within a document formalizes the notion of topic-sentiment informativeness. This is based on the Poisson distribution. Significant Theme words are identified using TF, Positional and Distribution factor. The distribution function for each theme word in a document is evaluated as follows (5.1):

$$f_d(w_i) = \sum_{i=1}^{n} (S_i - S_{i-1}) / n + \sum_{i=1}^{n} (TW_i - TW_{i-1}) / n \quad \text{----- (5.1)}$$

where n=number of sentences in a document with a particular theme word, S_i=sentence id of the current sentence containing the theme word and S_{i-1}=sentence id of the previous sentence containing the query term, TW_i is the positional id of current Theme word and TW_{i-1} is the positional id of the previous Theme word.

5.3.5.3.3 *Collocation*

Collocation with other thematic words/expressions is undoubtedly an important clue for identification of theme sequence patterns in a document. It has been observed that generally people does not change topic too frequently, i.e., a writer of a document does not switch too frequently

from subjective to objective rather he/she maintains a particular pattern. Based on these observations, the collocation feature has been introduced because topic-opinions are generally co-located. A window size of 5 including the present word is considered during training to capture the collocation with other thematic words/expressions.

5.3.6 Theme Clustering

Theme clustering algorithms partition a set of documents into finite number of topic based groups or clusters in terms of theme words/expressions. The task of document clustering is to create a reasonable set of clusters for a given set of documents. A reasonable cluster is defined as the one that maximizes the within-cluster document similarity and minimizes between-cluster similarities. There are two principal motivations for the use of this technique in theme clustering setting: efficiency, and the **cluster hypothesis**.

The **cluster hypothesis** (Jardine and van Rijsbergen, 1971) takes this argument a step further by asserting that retrieval from a clustered collection will not only be more efficient, but will in fact improve retrieval performance in terms of recall and precision. The basic notion behind this hypothesis is that by separating documents according to topic, relevant documents will be found together in the same cluster, and non-relevant documents will be avoided since they will reside in clusters that are not used for retrieval. Despite the plausibility of this hypothesis, there is only mixed experimental support for it. Results vary considerably based on the clustering algorithm and the document collection in use (Willett, 1988; Shaw et. al., 1996).

Application of the clustering technique to three sample documents results in the following theme-by-document matrix, A, where the rows represent Doc1, Doc7 and Doc13 (say) and the columns represent the themes politics, sport, and travel.

$$A = \begin{bmatrix} election & cricket & hotel \\ parliament & sachin & vacation \\ governor & soccer & tourist \end{bmatrix}$$

The similarity between vectors is calculated by assigning numerical weights to these words and then using the cosine similarity measure as specified in the following equation (5.2).

$$s\left(\vec{q}_k, \vec{d}_j\right) = \vec{q}_k . \vec{d}_j = \sum_{i=1}^{N} w_{i,k} \times w_{i,j} \text{ ----- (5.2)}$$

This equation specifies what is known as the dot product between vectors. Now, in general, the dot product between two vectors is not useful as a similarity metric, since it is too sensitive to the absolute magnitudes of the various dimensions. However, the dot product between vectors that have been length normalized has a useful and intuitive interpretation: it computes the **cosine** of the angle between the two vectors. When two documents are identical they will have a cosine of one; when they are orthogonal, i.e, they share no common terms and they will have a cosine of zero. If for some reason the vectors are not stored in a normalized form, then the normalization can be incorporated directly into the similarity measure as follows (5.3).

$$s\left(\vec{q}_k, \vec{d}_j\right) = \frac{\sum_{i=1}^{N} w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,k}^2}} \quad \text{----- (5.3)}$$

Of course, in situations where the document collection is relatively static, it makes sense to normalize the document vectors once and store them, rather than include the normalization in the similarity metric.

Calculating the similarity measure and using a predefined threshold value, documents are classified using standard bottom-up soft clustering k-means technique. The predefined threshold value is experimentally set to 0.5 as shown in the Table 5.6.

A set of initial cluster centers is defined in the beginning with the hypothesis that all words belong to a separate cluster. Therefore if total number of words is *k* then the initial cluster number is also *k*. Each document is assigned to the cluster whose center is closest to the document. After all documents have been assigned, the center of each cluster is recomputed as the centroid or mean $\vec{\mu}$ (where $\vec{\mu}$ is the clustering coefficient) of its members, that is $\vec{\mu} = \left(1/\left|c_j\right|\right)\sum_{x \in c_j} \vec{x}$. The distance function is the **cosine vector** similarity function.

ID	Themes	1	2	3
1	**প্রশাসন** (*administration*)	**0.63**	0.12	0.04
1	সুশাসন (*good-government*)	**0.58**	0.11	0.06
1	সমাজ (*Society*)	**0.58**	0.12	0.03
1	আইন (*Law*)	**0.55**	0.14	0.08
2	গবেষণা (*Research*)	0.11	**0.59**	0.02
2	কলেজ (*College*)	0.15	**0.55**	0.01
2	**উচ্চশিক্ষা (*Higher Study*)**	0.12	**0.66**	0.01
3	জেহাদি (*Jehadi*)	0.13	0.05	**0.58**
3	মসজিদ (*Mosque*)	0.05	0.01	**0.86**
3	মুশারফ (*Musharaf*)	0.05	0.01	**0.86**
3	**কাশ্মীর (*Kashmir*)**	0.03	0.01	**0.93**
3	পাকিস্তান (*Pakistan*)	0.06	0.02	**0.82**
3	নয়াদিল্লী (*New Delhi*)	0.12	0.04	**0.65**
3	বর্ডার (*Border*)	0.08	0.03	**0.79**

Table 5.6: Clustered Themes with Cluster Centroids (mean $\vec{\mu}_j$)

Table 5.6 gives an example of theme centroids of clusters based on K-means clustering. Bold words in Theme column are cluster centers. Cluster centers are assigned by maximum clustering coefficient. For each theme word, the cluster from Table 5.6 is still the dominating cluster. For example, "প্রশাসন" has a higher membership probability in cluster 1. But each theme word also has

some non-zero membership in other clusters. This is useful for assessing the strength of association between a theme word and a topic. Comparing two members of the cluster2, "কাশ্মীর" and "নয়াদিল্লী", it is seen that "নয়াদিল্লী" is strongly associated with cluster2 (p=0.65) but has some affinity with other clusters as well (e.g., p =0.12 with the cluster1). This is a good example of the utility of soft clustering. These non-zero values are still useful for calculating vertex weights during Theme Relational Graph generation.

5.3.7 Construction of Document Level Theme Relational Graph

Representation of input text document(s) in the form of a document graph is the key to our design principle. The idea is to build a document graph *G=<V,E>* from a given source document $d \in D$. First, the input document *d* is parsed and split into a number of text fragments (sentences) using sentence delimiters (Bengali sentence markers include "।", "?" or "!"). At this preprocessing stage, the text is tokenized, stop words are eliminated and words are stemmed. Thus, the text in each document is split into fragments and each fragment is represented with a vector of constituent theme words. These text fragments become the nodes *V* in the document graph.

The similarity between two nodes is expressed as the weight of each edge *E* of the document graph. A weighted edge is added to the document graph between two nodes if they either correspond to adjacent text fragments in the text or are semantically related by theme words. The weight of an edge denotes the degree of the relationship. The weighted edges not only denote document level similarity between nodes but also inter document level similarity between nodes. Thus to build a document graph *G*, only the edges with edge weight greater than some predefined threshold value are added to *G*.

The Cosine similarity measure has been used here. In cosine similarity, each document d is denoted by the vector $\vec{V}(d)$ derived from *d*, with each component in the vector is defined for each Theme word. The cosine similarity between two documents (nodes) d_1 and d_2 is computed using their vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$ as defined in equations 5.2 and 5.3. The dot product of two vectors $\vec{V}(d_1) \bullet \vec{V}(d_2)$ is defined as $\sum_{i=1}^{M} V(d_1)V(d_2)$. The Euclidean length of *d* is defined to be $\sqrt{\sum_{i=1}^{M} \vec{V}_i^{\,2}(d)}$ where M is the total number of documents in the corpus. Theme nodes within a cluster are connected by edges, whose weight is calculated by the clustering co-efficients of those theme nodes. No inter cluster vertices are present. Cluster centers are interconnected with weighted vertex. The weight is calculated by cluster distance based on cosine similarity measure discussed in section 5.3.6.

To better aid the understanding of the automatically determined category relationships the network has been visualized using the Fruchterman-Reingold force directed graph layout algorithm (Fruchterman and Reingold, 1991) and the NodeXL network analysis tool (Smith et. al., 2009)[13]. A constituent relational model graph drawn by NodeXL is shown in Fig. 5.22. In the graphical representation one color depict one cluster. The nodes of a cluster are connected with the cluster

[13] Available from http://www.codeplex.com/NodeXL

center. Relevance (in terms of network distance) between two element nodes could be calculated via cluster centers.

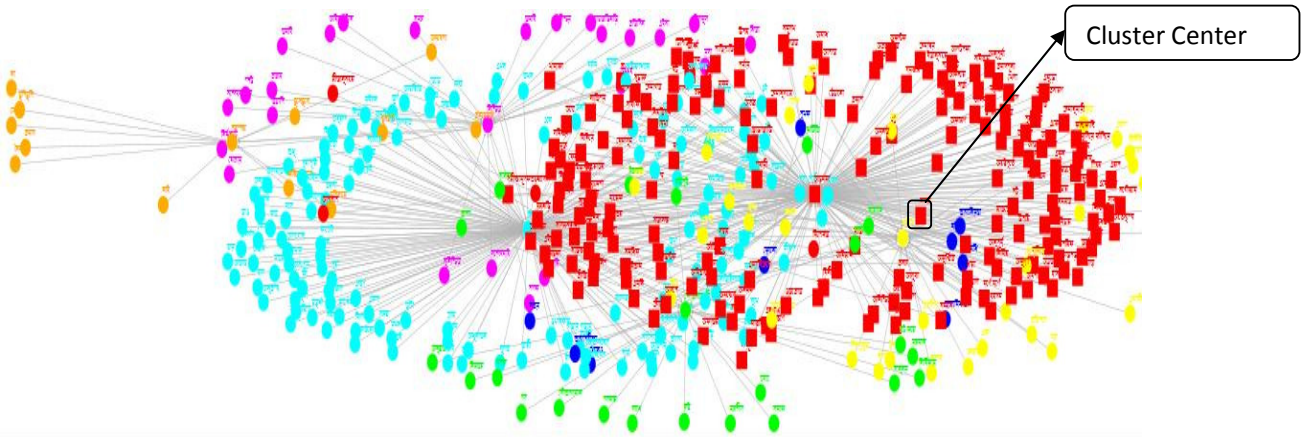


Figure 5.22: Document Level Theme Relational Graph by NodeXL

5.3.8 Summarization System

Candidate Sentence	Relevance Score
মহম্মদ আমিনের মতো **পলিটব্যুরোর** 'নবীনতম' সদস্যকেও কিন্তু বয়সের দিক হইতে নবীন ভাবা কঠিন।	151
এবার চিন্তা আরওএকটু বেশি, কারণ এই **মূল্যবৃদ্ধির** পিছনে যেমন দেশের ভিতরে জিনিসপত্রের জোগান কমে যাওয়া আছে, তেমনই আছে **আন্তর্জাতিক বাজারে মূল্যবৃদ্ধির** প্রবণতা।	167
স্বাধীনতার পর ষাট বছর গত হইল, এখনও প্রায় সকল **সরকারি পরিকল্পনার** পিছনে এই একটিই ভাবাদর্শ কাজ করে: বিভিন্ন **ভোটব্যাঙ্ককে** তুষ্ট করিয়া যেন তেন প্রকারেণ নিজেদের দলীয় স্থিতি নিশ্চিত করা।	130

Table 5.7: Candidate Sentences for Summary from Each Theme Cluster with the Relevance Scores

The present system is an extractive opinion summarization system for Bengali. In the section 5.3.6, the identification of theme clusters related to different shared topics and subtopics from a given input document set have been described. The next step is to extract thematic sentences from each theme cluster that reflects the contextual concise content of the current theme cluster. Extraction of sentences based on their importance in representing the shared subtopic (cluster) is an important issue and it defines the quality of the output summary. The popular Information Retrieval (IR) based Page-Rank algorithm has been used with slight modification to identify the most "informed" sentences from any cluster. With the adaptation of ideas from the Page-Rank algorithms (Page et. al., 1998), it can be easily observed that a text fragment (sentence) in a document is relevant if it is highly related to many relevant text fragments of other documents in the same cluster. Since, in our document graph structure, the edge score reflects the correlation measure between two nodes, it

can be used to identify the most salient/informed sentence from a sentence cluster. The relevance of a node/sentence is computed by summing up the edge scores of those edges that connect the node with other nodes in the same cluster. Then the nodes/sentences in each theme cluster are ranked according to their calculated relevance scores and the top ranked sentences in each theme cluster is selected as the candidate sentence representing the opinion summary. For example, some candidate sentences are shown in Table 5.7. The theme words are identified in bold. The sentences are extracted based on these theme words.

Once all the relevant sentences are extracted for each input document, these sentences are presented in the final summary in the original order in which they occurred in the original document.

5.3.9 Experimental Result of Multi-Document Topic-Opinion Extractive Summary

The evaluation result of the CRF-based Theme Detection task for Bengali is presented in Table 5.8. The result is presented individually on the annotated text for each annotator as well as the overall result of the system.

Theme Detection	Metrics	X	Y	Z	Avg.
	Precision	87.65%	85.06%	78.06%	83.60%
	Recall	80.78%	76.06%	72.46%	76.44%
	F-Score	84.07%	80.30%	75.16%	79.85%

Table 5.8: Performance of CRF-based Theme Identifier

Summarization	Metrics	X	Y	Z	Avg.
	Precision	77.65%	67.22%	71.57%	72.15%
	Recall	68.76%	64.53%	68.68%	67.32%
	F-Score	72.94%	65.85%	70.10%	69.65%

Table 5.9: Results of Subjective Sentence Identification for Opinion Summary

The evaluation results of subjective sentence identification of the system for opinion summary are shown in the Table 5.9. The results have been reported on the individual annotated text by each annotator as well as for the overall system.

5.3.10 Error Analysis

The evaluation result of the present summarization system is reasonably good but there are scopes of improvement. During the error analysis, it has been observed that the main reason lies with the subjectivity identifier. It has been reported in Chapter 2 that the recall value of the subjectivity

classifier is higher than its precision. Hence some objective sentences are identified during subjectivity analysis. Some of these sentences get high score during Theme detection or Theme clustering and are included in final summary.

Another vital source of errors occurs due to the accuracy level of linguistics tools, such as, POS tagger, Chunker and Dependency Parser. These linguistics tools do not have high accuracy figures and hence the Theme identification system misses some of the important theme words. Successive modules like Theme clustering and Document level weighted theme relational model fail to identify these important theme expressions.

5.4 The 5W Sentiment Summarization Visualization-Tracking

In today's digital age, text is the primary medium for representation and communication of information, as evidenced by the pervasiveness of e-mails, instant messages, documents, weblogs, news articles, homepages and printed materials. Our lives are now saturated with textual information and there is an increasing urgency to develop technology to help us manage and make sense of the resulting information overload. While expert systems have enjoyed some success in assisting information retrieval, data mining, and natural language processing (NLP) systems, there is a growing requirement of Sentiment Analysis (SA) systems that can automatically process the plethora of sentimental information available in online electronic text.

The Sentiment Analysis research has become quite matured after a few decades of activities in areas of sentiment knowledge acquisition, subjectivity detection, polarity classification and topic-opinion identification techniques. These related technologies have been discussed in earlier chapters in this thesis. The focus of this section is on aggregating and representing sentiment information drawn from an individual document or from a collection of documents. Sentiment/opinion aggregation is a necessary requirement at the end user's perspective. For example, an end user might desire an at-a-glance presentation of the main points made in a single review or might be interested to know how opinion changes with time over multiple documents. On real-life applications, the ultimate desired goal of the sentiment analysis research is to provide a completely automated solution. An intelligent system should be smart enough to aggregate all the scattered sentimental information from the various blogs, news article and from written reviews. The role of any automatic system is to minimize the effort of human users and generate a good acceptable output.

To provide fully automated summary or at-a-glance representation a system needs to know the semantic structure of a text. Structural opinion/sentiment analysis is one of the most important sub disciplines that need more attention to meet the user's need and satisfaction. Good structurization is required for in depth opinion/sentiment understanding. Philosophically speaking, opinion can be defined as the medium between knowledge and ignorance. But the question is: *what to know and what to ignore?* To answer this question, a relatively generic 5W structurization for opinions has been proposed and has been discussed in detail in Chapter 4. The 5W structured output has been used for further opinion summarization and visual tracking. The 5W task seeks to extract the semantic information of sentiment constituents in a natural language sentence by distilling it into

the answers to the 5W questions: **Who, What, When, Where** and **Why**. The visualization system facilitates users to generate sentiment tracking with textual summary and sentiment polarity wise graph based on any dimension or combination of dimensions as they want, i.e., "Who" are the actors and "What" are their sentiment regarding any topic, changes in sentiment during "When" and "Where" and the reasons for change in sentiment as "Why".

5.4.1 5W Constituent Clustering

The Theme clustering algorithms partition a set of documents into finite number of topic based groups or clusters in terms of their theme opinion constituents. The identified 5W constituents have been treated as a theme of those documents for the visualization and tracking work. The *k-means* soft clustering technique with appropriate modifications has been used for the topic-opinion clustering technique. The theme identification task produces a bag-of-words as the identified themes of the document. These themes are the 5W constituents and are classified into 5 distinct clusters. The modified clustering technique clusters documents based on these classes by calculating the similarity matrix by using the same formularization equations 5.2 and 5.3.

Application of the clustering technique to the three sample documents results in the following theme-by-document matrix, A, where the rows represent Doc1, Doc7 and Doc13 (say) and the columns represent the themes politics, sport and travel.

$$A \doteq \begin{array}{c} \begin{array}{ccccc} \text{Who} & \text{What} & \text{When} & \text{Where} & \text{Why} \end{array} \\ \begin{bmatrix} \text{Mamata Banerjee} & \text{Gyaneswari Express} & \text{24th May 2010} & \text{Jhargram} & \text{Maoist} \\ \text{West Bengal CM} & \text{Derailment} & \text{Midnight} & \text{Khemasoli} & \text{Bomb Blast} \\ \text{Pranob Mukherjee} & \text{Accident} & \text{Yesterday} & \text{Writers} & \text{Technical Fault} \end{bmatrix} \end{array}$$

Generated Clusters						
5Ws	**5W Opinion Constituents**	**Doc1**	**Doc2**	**Doc3**	**Doc4**	**Doc5**
Who	মমতা ব্যানার্জী (***Mamata Banerjee***)	**0.63**	0.01	**0.55**	**0.93**	0.02
	পশ্চিমবঙ্গের মুখ্যমন্ত্রী (*West Bengal CM*)	0.00	0.12	**0.37**	0.10	0.17
What	জ্ঞানেশ্বরী এক্সপ্রেস (***Gyaneswari Express***)	**0.98**	**0.79**	**0.58**	0.47	0.36
	লাইনচ্যুত (*Derailment*)	**0.98**	**0.76**	0.35	0.23	0.15
When	২৪ শে মে, ২০১০ (***24th May 2010***)	**0.94**	0.01	0.01	0.01	0.01
	মধ্যরাতে (*Midnight*)	**0.68**	**0.78**	0.01	0.01	0.01
Where	ঝাড়গ্রাম (***Jhargram***)	**0.76**	0.25	0.01	0.13	**0.76**
	খেমাসোলি (*Khemasoli*)	**0.87**	0.01	0.01	0.01	0.01
Why	মাওবাদী (***Maoist***)	**0.78**	**0.89**	0.06	0.10	0.14
	বিস্ফোরণ (*Bomb Blast*)	0.13	**0.78**	0.01	0.01	**0.78**

Table 5.10: Constituent Clustering by 5W Dimensions

An example output of the inter-document 5W constituent clustering technique has been reported in Table 5.10. The numeric scores are the similarity association values assigned by the clustering technique. Experimentally a threshold value of greater than 0.5 has been chosen. After the cluster

generation the document relational graph is built using the same method as discussed in the section 5.3.7.

5.4.2 Sentence Selection for Summary

In the generated constituent network all the lexicons are connected with weighted vertex either directly or indirectly. Semantic lexicon inference can be identified by network distance of any two constituent nodes by calculating the distance in terms of weighted vertex. The relevance of semantic lexicon nodes has been computed by summing up the edge scores of those edges connecting the node with other nodes in the same cluster. As cluster centers are also interconnected with weighted vertex so inter-cluster relations can be also calculated in terms of weighted network distance between two nodes within two separate clusters. Let us consider two clusters, *A* and *B* where *A* has *m* numbers of nodes while *B* consists of *n* numbers of nodes. a_x and b_y are the clusters centers of *A* and *B*.

$$A = \{a_1, a_2, a_3, a_4 ....., a_x, ...a_m\}$$

$$B = \{b_1, b_2, b_3, b_4 ....., b_y, ...b_n\}$$

The lexicon semantic affinity inference between a_x and b_y can be calculated as follows:

$$S_d(a_x, b_y) = \frac{\sum_{k=0}^{n} v_k}{k} \qquad ----- \ (5.4) \quad \text{or}$$

$$\sum_{c=0}^{m} \frac{\sum_{k=0}^{n} v_k}{k} \times \prod_{c=0}^{m} l_c \qquad ----- \ (5.5)$$

where $S_d(a_x, b_y) =$ semantic affinity distance between two constituent a_x and b_y, *k*=number of weighted vertex between two constituent a_x and b_y, v_k is the weighted vertex between two lexicons, *m*=number of cluster centers between two lexicons and l_c is the distance between cluster centers between two lexicons. Equations 5.4 and 5.5 are for intra-cluster and inter-cluster semantic distance measure respectively.

The present system is an extractive opinion summarization. The major step is to extract relevant sentences from each constituent cluster that reflects the contextual concise content of relevant cluster. The page rank algorithm first finds out the shortest distance which covers all the desired constituent nodes and maximizes the accumulated edge scores among them. Accordingly sentences are chosen which cover all the desired nodes. The working principle of the present system is as follows.

- The system identifies all the desired nodes in the developed semantic constituent network as given by user in the form of 5W.

- Inter-constituents distances have been calculated from the developed semantic constituent network. For example, based on the following input, the calculated inter-constituents distance values have been shown in Table 5.11.

Input:

Who	**What**	**When**	**Where**	**Why**
মমতা ব্যানার্জী	জ্ঞানেশ্বরী এক্সপ্রেস	মধ্যরাত	ঝাড়গ্রাম	মাওবাদী
(Mamata Banerjee)	(Gyaneswari Express)	(Midnight)	(Jhargram)	(Maoist)

Type	**Inter-Constituents Distances**				
	Who	What	When	Where	Why
Who	-	0.86	0.02	0.34	0.74
What	0.86	-	0.80	0.89	0.67
When	0.02	0.80	-	0.58	0.23
Where	0.34	0.89	0.58	-	0.20
Why	0.74	0.67	0.23	0.20	-

Table 5.11: Calculated inter-constituents distances for 5W Summarization-Visualization-Tracking

- All the sentences that consist of at least one of the user defined constituents are extracted from all the documents. The extracted sentences for the given user input are shown in Table 5.12.

- Extracted sentences are ranked with the adaptive Page-Rank algorithm based on the constituents present in that sentence. In the first iteration, the standard IR based Page-Rank algorithm assigns a score to each sentence based on keywords (constituents are treated as keywords in this stage) presence. In the second iteration, the calculated ranks by the Page-Rank algorithm are multiplied with the inter-constituents distances for those sentences where more than one constituent is present. For example, in the next sentence two Ws: "Who" and "What" are jointly present as constituent. Let us consider that the assigned rank for the following sentence by the basic Page-Rank algorithm is *n*. Then in the next iteration the modified score will be *n**0.86, because the inter-constituents distances for "Who" (মমতা বন্দ্যোপাধ্যায়) and "What" (জ্ঞানেশ্বরী এক্সপ্রেস) is 0.86.

  (মমতা বন্দ্যোপাধ্যায়)/**Who** (জ্ঞানেশ্বরী এক্সপ্রেস ঘটনাকে)/**What** রাজনৈতিক চক্রান্ত বলে মন্তব্য করেন।

  English Gloss: (Mamta Bandyopadhyay)/**Who** commented that the (Gyaneshwari Express incident)/**What** is a political conspiracy.

- The ranked sentences are then sorted in descending order and top-ranked 30% sentences (from all retrieved sentences) are shown as a summary.

Another issue that is very important in summarization is sentence ordering so that the output summary looks coherent. Once all the relevant sentences are extracted across the input documents, the summarizer has to decide in which order these sentences will be presented so that the whole text makes sense for the user. The temporal order of sentences as they occurred in original document has been preferred.

Extracted Sentences				
Who	**What**	**When**	**Where**	**Why**
মমতা ব্যানার্জী (Mamata Banerjee)	জ্ঞানেশ্বরী এক্সপ্রেস (Gyaneswari Express)	মধ্যরাতে (Midnight)	ঝাড়গ্রাম (Jhargram)	মাওবাদী (Maoist)
শ্রীমতী **মমতা বন্দ্যোপাধ্যায়** পরদিন সকালেই ঝাড়গ্রাম পৌছান ও প্রেস মিটিং-এ জানান, সিবিআইকে দিয়ে ঘটনাটি তদন্ত করা হচ্ছে।	এই দুর্ঘটনায় **জ্ঞানেশ্বরী এক্সপ্রেসের** লাইনচ্যুত বগিতে একটি চলন্ত মালগাড়ির ধাক্কা লাগায় মোট ১৪১ জন যাত্রীর প্রাণহানি ঘটে।	প্রায় **মধ্যরাতে** উল্টোদিক থেকে আসা একটি মালগাড়ি লাইনচ্যুত ট্রেনটিতে ধাক্কা মারে এবং এই দুর্ঘটনায় ১৪১ জনের মৃত্যু হয় এবং ১৮০ জনেরও বেশি আহত হন।।	**ঝাড়গ্রামের** কাছে খেমাশোলি ও সরডিয়া স্টেশনের মধ্যবর্তী অংশে এই দুর্ঘটনা ঘটে	এই ট্রেন দুর্ঘটনার কয়েক সপ্তাহ আগে রেলমন্ত্রী মমতা বন্দ্যোপাধ্যায় বলেন যে **মাওবাদী** হানার ফলে ভারতীয় রেল আর্থিক ক্ষতির সম্মুখীন হচ্ছে।
তদন্ত শুরু করেছে সিআইডি, তবে রেলমন্ত্রী মমতা বন্দ্যোপাধ্যায় ট্রেন বেলাইন হওয়ার কারণ হিসেবে রেললাইনে বিস্ফোরণ ঘটার তথ্য দিয়েছেন, যার কোনও প্রমাণ পাওয়া যায়নি।	হাওড়া-কুরলা লোকমান্য তিলক জ্ঞানেশ্বরী সুপার ডিলাক্স এক্সপ্রেস (সংক্ষেপে "**জ্ঞানেশ্বরী এক্সপ্রেস**") নামে এই ট্রেনটি হাওড়া থেকে মুম্বই যাচ্ছিল।	.....	.....	ভারতের রেলমন্ত্রী মমতা বন্দ্যোপাধ্যায় বলেছেন, **মাওবাদীদের** দ্বারা ঘটানো শক্তিশালী বোমা বিস্ফোরণেই ট্রেনটি লাইনচ্যুত হয়েছে।
এমনকী এই ঘটনা যে পুরভোটের আগে তাঁকে বেকায়দায় ফেলার চক্রান্ত, এমন ইঙ্গিতও দিয়েছেন মমতা বন্দ্যোপাধ্যায়।	.....	.....	.....	রেলমন্ত্রীর বক্তব্য, **মাওবাদীরা** বোমা বিস্ফোরণ ঘটিয়ে "পরিকল্পিত হামলা" চালানো হয়, যার ফলে ট্রেনটি লাইনচ্যুত হয়।
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....

Table 5.12: Extracted Sentences for 5W Summarization-Visualization-Tracking

5.4.3 Dimension Wise Opinion Summary, Visualization and Tracking

The visualization system consists of five drop down boxes. The drop down boxes give options for individual 5W dimension of each unique W that is present in the corpus. The present visualization system facilitates users to generate opinion summary and opinion polarity wise graph based visualization and summary on any 5W dimension or combination of 5W dimensions as they want. The present system also provides an overall summary and visualization. To aggregate the sentimental information at any direction of the proposed 5W constituents, the same *k-means* clustering technique has been used.

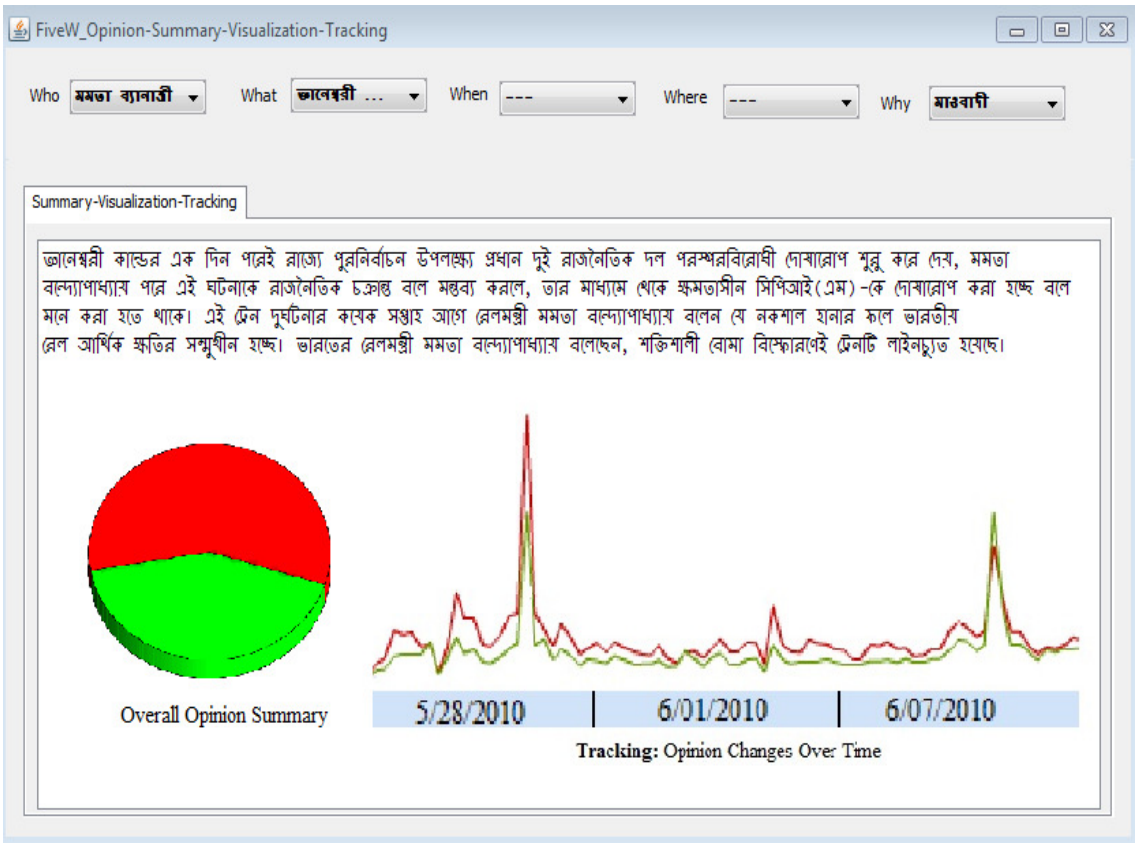


Figure 5.23: A Snapshot of the 5W Summarization-Visualization-Tracking System

To generate dimension specific and overall polarity wise graph the polarity scores assigned by the polarity identifier (described in Chapter 2) have been used. The page rank algorithm (Page et. al., 1998) has been adopted to generate dimension specific summaries. The Page-Rank algorithm searches for the shortest path within the developed network and attempts to cover all the input nodes by the user in the best possible way. The inter-document theme relational graph has been used to identify the shared subtopic clusters in terms of the shared 5W opinion constituents. The extracted sentences are then sorted according to their associated scores obtained by summing up the edge scores of those edges connecting the 5W node with other 5W nodes in the same cluster. Top ranked 30% of total extracted sentences are chosen as a dimension wise summary. But the graph is built using the scores from the extracted sentences and ordered time wise based on the "When" tag. Dimension specific summary and visualization gives an in depth understanding of any kind of opinion/sentiment or emotion over any domain or language. A snapshot of the present system has been shown in Figure 5.23.

Another important aspect of the present system is that a user can provide no input along any direction to see the all the possible information on that direction. Some example outputs from the present system are shown below.

Example 1

Input:	**Who**	**What**	**When**	**Where**	**Why**
	মমতা ব্যানার্জী	লাইনচ্যুত	---	ঝাড়গ্রাম	---

English Gloss: Mamata Banerjee Derailment --- Jhargram ---

Textual Summary: পরশু মধ্যরাতে ঝাড়গ্রামের অদূরে জ্ঞানেশ্বরী এক্সপ্রেসের লাইনচ্যুত হওয়ার ঘটনাকে বড়সড় রাজনৈতিক ষড়যন্ত্র বলে দাবি করেন মমতা। শ্রীমতী মমতা বন্দ্যোপাধ্যায় পরদিন সকালেই ঝাড়গ্রাম পৌছান ও প্রেস মিটিং-এ জানান, সিবিআইকে দিয়ে ঘটনাটি তদন্ত করা হচ্ছে। তদন্ত শুরু করেছে সিআইডি, তবে রেলমন্ত্রী মমতা বন্দ্যোপাধ্যায় ট্রেন বেলাইন হওয়ার কারণ হিসেবে রেললাইনে বিস্ফোরণ ঘটার তথ্য দিয়েছেন, যার কোনও প্রমাণ পাওয়া যায়নি। এমনকী এই ঘটনা যে পুরভোটের আগে তাঁকে বেকায়দায় ফেলার চক্রান্ত, এমন ইঙ্গিতও দিয়েছেন মমতা বন্দ্যোপাধ্যায়।

English Gloss: Mamta claimed that the derailment incident of the Jyaneswari Express near Jharagramera, happened day before tomorrow at midnight is a big political conspiracy. Smt. Mamta Bandyopadhyay reached Jharagrama next morning and said in press meeting event that the case will be investigated by CBI. CID has started investigation, but rail minister Mamta Bandyopadhyay has given a theory of explosion as a probable reason of derailment of the train, of which no evidence has been still found. Even this inci-dent before the municipality election is a conspiracy to make her defeat, Mamta Bandyopadhyay has given this indications.

Example 2

Input:	**Who**	**What**	**When**	**Where**	**Why**
	মমতা ব্যানার্জী	জ্ঞানেশ্বরী এক্সপ্রেস	---	---	মাওবাদী

English Gloss: Mamata Banerjee --- --- --- Maoist

Textual Summary: জ্ঞানেশ্বরী কান্ডের এক দিন পরেই রাজ্যে পুরনির্বাচন উপলক্ষ্যে প্রধান দুই রাজনৈতিক দল পরস্পরবিরোধী দোষারোপ শুরু করে দেয়, মমতা বন্দ্যোপাধ্যায় পরে এই ঘটনাকে রাজনৈতিক চক্রান্ত বলে মন্তব্য করলে, তার মাধ্যমে থেকে ক্ষমতাসীন সিপিআই(এম)-কে দোষারোপ করা হচ্ছে বলে মনে করা হতে থাকে। এই ট্রেন দুর্ঘটনার কয়েক সপ্তাহ আগে রেলমন্ত্রী মমতা বন্দ্যোপাধ্যায় বলেন যে নকশাল হানার ফলে ভারতীয় রেল আর্থিক ক্ষতির সম্মুখীন হচ্ছে। ভারতের রেলমন্ত্রী মমতা বন্দ্যোপাধ্যায় বলেছেন, শক্তিশালী বোমা বিস্ফোরণেই ট্রেনটি লাইনচ্যুত হয়েছে।

English Gloss: One day after the Jyaneswari incident, due to Municipality election in the State two main political parties began to blame conflicting each other and Mamta Bandyopadhyay commented the incident as a political conspiracy, through that it is considered that the ruling CPI (M) party is being blamed. A few weeks before the train accident rail minister Mamta Bandyopadhyay

said that due to attack by the nakasala the Indian Railways is experiencing financial losses. The rail minister Mamta bandyopadhyaya said, the train has derailed due to explosion of powerful bomb.

The previous two examples show that the proposed system attempts to satisfy the information need of the user. In the example 1, the user wants to know the explanations of the probable causes by *Mamata Banerjee* for the *train accident*. But in the example 2, the user wants to find out *Mamata Banerjee's* comment on the *Maoists* involvement in the *Gyaneswari* train accident.

5.4.4 Experimental Result of the 5W Sentiment Summarization-Visualization-Tracking

To evaluate the present system a two-fold evaluation mechanism has been followed. The first-fold evaluation is to understand the system performance to detect subjective sentences prior to generation of the final summary (as mentioned in the third step of the Summary process). The system identifies all the sentences that consist of at least one of the user defined constituents. These sentences are extracted from all the documents. For evaluation, the system identified sentences are checked with every human annotator's gold standard sentences and finally the overall accuracy of the system is calculated as reported in Table 5.13.

Summarization	Metrics	X	Y	Z	Avg.
	Precision	77.65%	67.22%	71.57%	72.15%
	Recall	68.76%	64.53%	68.68%	67.32%
	F-Score	72.94%	65.85%	70.10%	69.65%

Table 5.13: Evaluation Results of the Summarization System

Numeric Scores	Measure
1	Very Poor
2	Poor
3	Acceptable
4	Good
5	Excellent

Table 5.14: 5-point Scoring Standards for Summary Evaluation

It was a challenge to evaluate the accuracy of the dimension specific summaries. According to the classical theory, human developed gold standard dataset should have been prepared for every dimension combinations, but it is too difficult to develop such gold standard dataset. A direct

human evaluation technique has been proposed in the present work. Two evaluators have been involved in the present task and they have been asked to give score to each system generated summaries. The 5-point scoring technique used in the present system has been shown in Table 5.14. The final evaluation result of the dimension specific summarization system is reported in Table 5.15.

Tags	Average Scores				
Who	What	When	Where	Why	Overall
	3.20	3.30	3.30	2.50	3.08
What	Who	When	Where	Why	Overall
	3.20	3.33	3.80	2.6	3.23
When	Who	What	Where	Why	Overall
	3.30	3.33	2.0	2.5	3.00
Where	Who	What	When	Why	Overall
	3.30	3.80	2.0	2.0	2.77
Why	Who	What	When	Where	Overall
	2.50	2.6	2.5	2.0	2.40

Table 5.15: Subjective Human Evaluation Results on 5W Dimension Specific Summaries

Publications

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010. ***Opinion Summarization in Bengali: A Theme Network Model***. *In the Proceeding of the 2nd IEEE International Conference on Social Computing (SocialCom-2010)*, Pages 675-682, Minneapolis, USA. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5591520

2. **Amitava Das** and Sivaji Bandyopadhyay. ***Topic-Based Bengali Opinion Summarization***. 2010. *In the Proceeding of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Pages 232-240, Beijing, China. http://aclweb.org/anthology/C/C10/C10-2027.pdf

3. **Amitava Das** and Sivaji Bandyopadhyay. 2012(d). ***The 5W Structure for Sentiment Summarization-Visualization-Tracking***. *In the Proceeding of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, Delhi, India. (**Accepted**).

Conclusion

An account of the key scientific contributions of this thesis along with a brief roadmap of the future possible avenues of this work has been reported in this chapter. Besides the manual development of linguistics resources and supplementary natural language processing tools like stemmer and dependency parser for Bengali, the thesis makes key scientific research contributions in various areas of sentiment analysis that includes Sentiment Lexicon Acquisition, Sentiment / Subjectivity Detection, Sentiment Polarity Detection, Sentiment Structurization and Sentiment Summarization-Visualization-Tracking.

Sentiment analysis in natural language text is a multifaceted and multidisciplinary Artificial Intelligence (AI) problem. It tries to narrow the communicative gap between the highly sentimental human and the sentimentally challenged computers by developing computational systems that recognize and respond to the sentimental states of the human users. There is a perpetual debate about the better way of collecting intelligence either by following the functional path of biological human intelligence or generating new methodologies for completely heterogeneous mechatronics machine that redefine a completely new horizon called *electronic intelligence*. Actually, we need an optimized solution between the biological human intelligence and the desired electronic intelligence. In this context, the following comment by **Professor Eduard Hovy** during the keynote lecture at the International Joint Conference on Natural Language (IJCNLP) 2011 Workshop '*Sentiment Analysis where AI Meets Psychology (SAAIP)*[1]' looks appropriate.

"*Today's airplane does not flutter its wings like birds but still it is relatively more capable than its ideological father, i.e. birds to fly more distance and to carry huge weight.* "

With this philosophy in mind the research endeavor in the present work was to find out the optimum solution strategies for computers that can either mimic the techniques of self-organized biological human intelligence or can at least simulate the functional similarities of human sentimental intelligence.

C.1 Contribution: ***Sentiment Knowledge Acquisition***

Sentiment knowledge acquisition in terms of sentiment lexicon is the vital pre-requisite of any sentiment analysis system. Previous studies have proposed to attach **prior polarity** scores to each sentiment lexicon. Prior polarity values are approximations obtained from corpus heuristics.

A number of research endeavors can be found in the literature for creation of sentiment lexicons in several languages and domains. These techniques can be broadly categorized into two genres, one follows the classical manual annotation techniques and the other proposes various automatic techniques. Both types of techniques have few limitations. Manual annotation techniques are undoubtedly trustable but it generally takes time. Such techniques require a large number of annotators to balance the sentimentality of individual annotators in order to reach agreement. But human

[1] http://saaip.org/

annotators are quite unavailable and costly. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain.

Both the processes have been attempted in the present work to develop SentiWordNet(s) (Das and Bandyopadhyay, 2010(d)) for multiple languages. During evaluation it was noticed that there are two issues that should be satisfied by a good quality sentiment lexicon: the first one is **coverage** and the second one is **credibility** of the associative polarity score. At the end of the journey, it may be concluded that automatic processes are good for coverage expansion but manual methods are trustable for prior polarity assignment.

The following automatic processes have been used: ***bilingual dictionary based, WordNet based synonym and antonym expansion, orthographic antonym generation*** and ***monolingual corpus based approach***. English sentiment lexicon is chosen as the source and the synset members are translated into the target language using bilingual lexicons. WordNet(s) have been used effectively to expand the synsets via synonym or antonym search. Sixteen hand crafted suffix/affix rules have been used to orthographically create more antonyms for a given synset. These antonyms have been confirmed using corpus validation techniques. The generated sentiment lexicon is used as a seed list. Language specific corpus is then automatically tagged with these seed words using the simple tagset of SWP (Sentiment Word Positive) and SWN (Sentiment Word Negative). A Conditional Random Field (CRF) based classifier has been trained on this tagged corpus. The CRF based system has been used on un-annotated corpus to find out new sentimental words. These techniques have been successfully used for three Indian languages: **Bengali, Hindi** and **Telugu** (Das and Bandyopadhyay, 2010(c); Das and Bandyopadhyay, 2010(e)). **Bengali SentiWordNet**[2] has been successfully developed and made available for further research.

Automatic processes are helpful to increase the coverage of the developed SentiWordNet(s) but human annotators make the resources credible. Sentiment is not a property of language. No linguistics theory can answer which property of a word gives a clue that the word depicts positive or negative sentiment. Therefore only a human annotator can help to get perfect credible polarity scores. But there is acute scarcity of human annotators; it was decided to involve **Internet Population** for creating more credible sentiment lexicons (Das and Bandyopadhyay, 2011). Internet population is very huge in number and ever growing. According to some estimates, there are approximately 360,985,492[3] numbers of people with various languages, cultures, age etc. and thus is not biased towards any domain, language or particular society. An online game called **Dr Sentiment** has been developed which is a template based interactive online game. **Dr Sentiment** collects the sentiments of the player by asking a set of simple template based questions and finally reveals the sentimental status of the player. The lexicons tagged by this system are credible as it is tagged by human beings. In either way it is not like a static sentiment

[2] http://www.amitavadas.com/sentiwordnet.php
[3] http://www.internetworldstats.com/stats.htm

lexicon set as its prior polarity scores are updated regularly. Almost 100 players per day are currently playing it throughout the world in different languages. Based on the immense success of this method the **Global SentiWordNet** (Das and Bandyopadhyay, 2010(d)), the SentiWordNet(s) for **57 languages,** has been developed using Google translation API[4] services.

Moreover, the lexicon generated and validated by the internet population using Dr Sentiment is equipped with psychological information on a few aspects at present along with sentiment knowledge. The sentiment lexicon wrapped with sentiment and psychological information together is called **PsychoSentiWordNet** (Das, 2011). The PsychoSentiWordNet helps to capture the overall picture of human social psychology regarding sentiment understanding. The changes of sentimentality with age, gender and geo-spatial location have been studied in the present work.

C.1.1 Points of Contribution on Sentiment Knowledge Acquisition

- Automatic processes are good for coverage expansion.
- Manual methods are credible for prior polarity assignment.
- Human annotators are necessary as sentiment is not the property of language.
- Bengali SentiWordNet has been successfully developed and made available for further research.
- SentiWordNet has been developed for two other Indian languages, Hindi and Telugu.
- Global SentiWordNet has been developed for 57 languages.
- The idea of PsychoSentiWordNet has been introduced.

C.1.2 The Road Ahead: ***Sentiment Knowledge Acquisition***

The Bengali SentiWordNet developed as part of the present work has been made available to the research community to carry on further research. As SentiWordNet is being developed for several other languages, there is an immediate research potential on cross-lingual sentiment sense mapping. Assignment of sense ID to each sysnset is necessary to start this task and this work has already been initiated.

The motivation behind the development of the PsychoSentiWordNet is to generate up-to-date prior polarity scores across various dimensions. The future goal is to generate web service APIs through which the research community can access the latest prior polarity scores for sentiment analysis research.

[4] http://translate.google.com/

C.2 Contribution: ***Sentiment / Subjectivity Detection***

The term subjectivity refers to the identification of Topical Relevant Sentiment in a piece of text. Sentiment/ subjectivity detection is a very tough challenge for emotionally challenged machines.

A series of experiments has been carried out to identify the optimum, easily extractable and generic feature set for English and Bengali language. On the contrary, the previous research efforts concentrated only on a few domain dependent lexical and syntax level features. The identified features are classified into three genres, Lexico-Syntactic, Syntactic and Discourse level features.

Experiments on Sentiment / Subjectivity detection have been carried out using **Rule-based, Machine Learning** and **Hybrid** techniques. But the present NLP techniques are inadequate to identify the clues related to human psychology or inter-relations of the linguistic clues that play a major role in the Sentiment / Subjectivity detection task. Experiments with **Genetic Algorithm** started next to adopt the biological evolutionary path of the human intelligence for machines. Interestingly, the Genetic Based Machine Learning (GBML) technique increases the accuracy of the system to **90.22%** and **90.6%** for English (on the MPQA corpus) and Bengali (on Blog corpus) respectively. This is the highest performing sentiment / subjectivity detection score among the whole literature till date.

Application of machine learning algorithms in natural language processing (NLP) generally experiments with combination of various syntactic and semantic linguistic features to identify the most effective feature set. In the present work, the sentiment / subjectivity detection problem has been viewed as a **Multi-Objective** or **Multi-Criteria Optimization** search problem. The experiments started with a large set of extractable syntactic, semantic and discourse level feature set. The **fitness function** calculates the accuracy of the subjectivity classifier based on the feature set identified by **natural selection** through the process of **crossover** and **mutation** after each generation. The GBML technique automatically identifies the best feature set based on the principle of natural selection and **survival of the fittest**. The identified fittest feature set is optimized **locally** and **global optimization** is obtained by multi-objective optimization technique.

C.2.1 Points of Contribution on Sentiment / Subjectivity Detection

- This is the first attempt on Bengali sentiment / subjectivity detection.
- The optimum, easily extractable, generic and classified feature set has been identified.
- The Multiple Objective Optimization through Genetic Algorithm has been shown as a good success for the sentiment / subjectivity detection task.
- The sentiment / subjectivity detection system based on Genetic Algorithm (GA) is the highest performing system till date for English and Bengali.

C.2.2 The Road Ahead: **Sentiment / Subjectivity Detection**

The genetic algorithm (GA) based technique generates huge success for the sentiment / subjectivity detection task. Yet, it cannot be claimed that the technique can mimic human biological mechanism to produce more intelligent machine. Presently, experiments have been initiated with other machine learning algorithms like artificial neural network which are fundamentally inspired from human intelligence theory. Such experiments may reveal some unidentified clues of human sentiment understanding.

C.3 Contribution: ***Sentiment Polarity Detection***

The polarity classification task involves sentiment/opinion classification into semantic classes such as *positive, negative or neutral* and/or other fine-grained emotion classes like *happy, sad, anger, disgust and surprise* etc. Various polarity classification techniques have been proposed in the present work.

The two step methodology, i.e., use of prior polarity lexicon followed by any NLP technique is the standard method for the polarity classification task as established by several previous research efforts. The SentiWordNet (Bengali) has been developed as part of the present work as the prior polarity lexicon. For the NLP technique, the experiments started with the Statistical Syntactic classification technique. The syntactic clue directly helps to understand the relation between the localized semantic orientation, i.e., word level semantic orientation and the contextual semantic orientation, i.e., word/phrase/sentence level semantic orientation. The Support Vector Machine (SVM) has been used with a number of features for the development of the syntactic polarity classifier. The polarity classification task mainly involves **syntactic analysis** with features like **Modifier-Modified** relationship or **Association ambiguity**. Therefore, the development of a **dependency parser** for Bengali language was attempted as there was no dependency parser available for Bengali. Dependency parsing properly analyzes the syntactic structure of a language. The details of the Dependency Parser development are described in the Appendix. Many other linguistics features like negative word, stemming cluster, functional word, part of speech and chunk information have been included in the polarity classifier. It has been established through **Feature ablation** method that the SentiWordNet based polarity classifier provides a good **baseline** and the best accuracy score of 70.04% is achieved with other features i.e. negative word, stemming cluster, functional word, part of speech, chunk and dependency tree features.

Dealing with unknown/new words is a common problem for NLP systems. It becomes more difficult for sentiment analysis because it is very hard to find out any contextual clue to predict the sentimental orientation on any unknown/new word. A prior polarity lexicon is attached with two probabilistic values, i.e., positivity and negativity scores but there is no clue in the SentiWordNet regarding ***which value to pick in which context?*** . The general trend is to pick the polarity corresponding to the highest score but that may vary depending on the context.

The research attempts in the present work are mainly concerned with the handling of the ambiguous entries in the SentiWordNet where the positive and negative scores are both non zero. The basic hypothesis is that addition of some contextual information along with the prior polarity scores in the sentiment lexicon lessens the requirement of further NLP techniques to disambiguate the contextual polarity. A new paradigm called ***Sentimantics*** has been introduced in the present work which is a distributed semantic lexical model to hold the sentiment knowledge with contextual common sense.

Two different types of models for Sentimantic composition have been examined that are empirically grounded and can represent the contextual similarity relations among various lexical sentiment and non-sentiment concepts. The Vector Space Model (VSM) for Sentimantics starts with existing resources like ConceptNet and SentiWordNet for English and SemanticNet and SentiWordNet (Bengali) for Bengali. The common sense lexicons like ConceptNet and SemanticNet are developed for general purpose use and the formalization of Sentimantics from these resources suffers due to lack of dimensionality. Therefore the Vector Space Model (VSM) has been developed to hold the Sentimantic from scratch by a corpus driven semi-supervised method. Generally, extracting knowledge from this kind of VSM is algorithmically very expensive because it is a very high dimensional network. Another important limitation of this type of model is that it demands very well defined processed input, like, Input: (high) Context (sensex, share market, point), which demands a NLP pre-processing on the input text to extract knowledge from this VSM. Finally, the Syntactic Co-occurrence Based VSM with relatively fewer dimensions has been proposed. The final model is the best performing lexicon network model which can be described as the acceptable solution for the Sentimantics problem. Each sentiment word in the developed lexical network is assigned a contextual prior polarity. The details of the proposed models are described in the Chapter three.

C.3.1 Points of Contribution on Sentiment Polarity Detection

- Syntactic relations help to detect sentiment polarity immensely at sentence level.
- Importance of the syntactic features has been established through feature ablation method.
- The idea of Sentimantics has been introduced.

C.3.2 The Road Ahead: **Sentiment Polarity Detection**

In future there is immense possibility of exploring the Sentimantics concept. Another possible work in future is to merge the two proposed lexical resources i.e., the **PsychoSentiWordNet** and **Sentimantics** together for more vibrant sentiment knowledge representation.

C.4 Contribution: ***Sentiment Structurization***

The need of the end user is the driving force behind the sentiment analysis research. Therefore, the outcomes of the sentiment analysis research should lead to the development of a real time sentiment analysis system which will successfully satisfy the need of the end users. The end users not only look for the binary (positive/negative) sentiment classification of an entity but they also become interested in aspectual sentiment analysis, identification of sentiments on different aspects of an entity. A sentiment analysis system should be capable enough to understand and extract the aspectual sentiments present in a natural language text.

Previous research efforts have already proposed various structures or components for sentiment extraction. Among the proposed sentiment structures, the widely used one includes **Holder**, **Topic** and other domain dependant **Attributes**. But the real life users do not always require all the aspects at a time, rather they look for opinion/sentiment changes of any "Who" during "When" and depending upon "What" or "Where" and "Why". With this hypothesis the 5W (**Who**, **What**, **When**, **Where** and **Why**) constituent extraction technique has been proposed for sentiment/opinion structurization. The proposed 5W structure is domain independent and more generic than the existing semantic constituent extraction structures. The proposed structure is also in agreement with the holder, topic or attributes model as "Who" represents the holder, "What" represents the topic and the other 3Ws represent the necessary attributes.

All the 5W constituents do not occur regularly in the corpus. Hence, the sequence labeling task with 5Ws tags using any machine learning technique will lead to a label bias problem and thus may not be an acceptable solution for the 5W role labeling task. A 5W role labeling task has been proposed in the present task that follows hybrid architecture. The system first assigns 5W labels to each chunk in a sentence using the Maximum Entropy Model (MEMM) and then a rule based post-processor helps to reduce many false hits by the MEMM based system. The rule-based model also identifies new 5W labels. The rules have been developed based on the acquired statistics on the training set and the linguistic analysis of standard Bengali grammar. It has been established that the hybrid structure is essential for the 5W role labeling task.

C.4.1 Points of Contribution on Sentiment Structrization

- The 5W (Who, What, When, Where and Why) structure has been introduced
- The 5W structure is more generic and the acceptable solution across domains.

C.4.2 The Road Ahead: **Sentiment Structurization**

The proposed 5W structure is a domain independent generic structure for the sentiment analysis task. The 5W structure is also helpful for Event or Sentiment-Event tracking tasks. Relevant experiments have been initiated in this direction.

C.5 Contribution: ***Sentiment Summarization-Visualization-Tracking***

Presentation of information to the end user in an aggregated format (summarization, visualization, and tracking) is the necessity from the end users' perspective but it is nearly impossible to develop a consensus on the output format or how the data should be aggregated. Researchers have attempted with various types of output formats like textual or visual summary or overall tracking along the time dimension. There are several research attempts on **Topic-wise** and **Polarity-wise** summarization and on **Visualization** and **Tracking**. Actually the output format varies on the requirements of the end user as well as on the domain. In the present work, multiple output formats have been experimented.

The experiments started with the multi-document topic-opinion textual summary but realizing the end user's requirements and the need to present an *at-a-glance* presentation, the 5W constituent based textual summarization-visualization-tracking system has been devised. The 5W constituent based aggregation system is a multi-genre system. The system facilitates users to generate sentiment tracking with textual summary and sentiment polarity wise graph based on any dimension or combination of dimensions as they want. To the best of our knowledge, the 5W constituent based summarization-visualization-tracking system attempts to answer the philosophical question of "*Topic-Wise*, *Polarity-Wise* or *Other-Wise*" raised by Dasgupta and Ng (Dasgupta and Ng, 2009).

Topic-Wise: The system facilitates users to generate sentiment summary based on any customized topic like Who, What, When, Where and Why based on any dimension or combination of dimensions as they want.

Polarity-Wise: The system produces an overall gnat chart that can be treated as the overall polarity wise summary. An interested user can still look into the summary text to find out more details.

Visualization and Tracking: The system facilitates users to generate visual sentiment tracking with polarity wise graph based on any dimension or combination of dimensions as they want, i.e., "Who" are the actors and "What" are their sentiment regarding any topic, changes in sentiment during "When" and "Where" and the reasons for change in sentiment as "Why". The final graph for the tracking is generated with a timeline. Moreover the end user can structure their information need as:

- ***Who***? Who was involved?
- ***What***? What happened?
- ***When***? When did it take place?
- ***Where***? Where did it take place?
- ***Why***? Why did it happen?

During the development of the multi-document topic-opinion summarization system, a strong semantic lexical network has been proposed following the idea of Mental Lexicon models. The same lexical semantic network has been used to develop the 5W summarization-visualization-tracking system as well.

C.5.1 Points of Contribution on Sentiment Summarization-Visualization-Tracking

- Distributional Semantics or Mental Lexicon Model is very effective to solve any kind of semantic inference problem. A similar model has been used for Topic-Opinion Summarization.
- The 5W Summarization-Visualization-Tracking is an acceptable solution for satisfying the end users' requirements across domains.

C.5.2 The Road Ahead: *Sentiment* Summarization-Visualization-Tracking

The proposed 5W model can be effectively used in Sentiment search or Sentiment translation. Experiments have started in this direction. The 5W model may also be useful for Event-Sentiment Tracking.

Appendix

Stemmer and Dependency Parser

Development of resources and tools are one of the most challenging tasks while working with resource constrained languages like Bengali. Bengali is the fifth popular language in the World, second in India and the national language of Bangladesh. Extensive Natural Language Processing (NLP) research activities in Bengali have started recently but annotated corpus and various linguistic tools are still unavailable for Bengali in the required measure. Corpus developments for subjectivity, polarity, structurization and summarization-visualization-tracking tasks have already been discussed in the respective chapters. In this Appendix section, the development of two main NLP tools, i.e., **Stemmer** and **Dependency Parser** are discussed. These tools are necessary to pursue any NLP research in Bengali.

A.1 Cluster based Stemming for Bengali

A Morphological Parser of Bengali (ethnonym: Bangla; exonym: Bengali) words using stemming cluster technique have been developed as part of the present work. The addition of inflectional suffixes, derivational suffixes and agglutination in compound words make Morphological Parsing fairly complex for the Bengali. Only one research attempt could be found at building a complete morphological parser for Bengali (Dasgupta and Khan, 2004). But unfortunately the software is not publicly available. Therefore the primary necessity was to develop a morphological parser or at least a quickly developed stemmer that can extract the root word/stem from an inflected surface word.

Morphological Parsing in Information Retrieval aspect does not demand identification of full morphological feature structure always. Identification of stems from several surface forms of a particular word is required. Highly motivated by the success of Porter Stemmer for English, a Morphological stemmer based on stemming cluster technique has been developed for Bengali.

The present stemmer analyzes prefix and suffix features of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a cluster with the identified root word as the cluster centre. An inflectional suffix is a terminal affix that does not change the word-class (parts of speech) of the root during concatenation; it is added to maintain the syntactic environment of the root in Bengali. On the other hand, derivational suffixes change word-class (parts of speech) and the orthographic form of the root word.

Experiments have been carried out with two types of algorithms: simple suffix stripping algorithm and score based stemming cluster identification algorithm. The Suffix stripping algorithm simply checks if any word has any suffixes (one suffix or more than one) in a manually created suffix list. The word is then assigned to the appropriate cluster whose cluster centre is the assumed root word, i.e., the form obtained after deleting the suffix from the surface form. Suffix stripping algorithm works well for Noun, Adjective and Adverb categories. The words of other part of speech categories, especially Verbs follow derivational morphology. The score based stemming technique has been designed to resolve the stem for inflected word forms. The technique uses Minimum Edit Distance method (Kukich, 1992), well known for spelling error detection, to measure the cost of identification of the class of every word. Score based

technique considers two standard operations of Minimum Edit Distance, i.e., insertion and deletion. Insertion and deletion of up to three characters have been considered in the present work. The idea is that the present word matches an existing cluster centre after insertion and/or deletion of maximum three characters. The present word will be assigned to the cluster that can be reached with minimum number of insertion and/or deletion. This is an iterative clustering mechanism for assigning each word into a cluster. A separate list of verb inflections (only 50 entries; manually edited) has been maintained to validate the result of the score based technique. The standard K-means Clustering technique has been used here. Each cluster centre is treated as a root stem. The system has reported an accuracy of 74.6%.

A.1.2 Previous Studies on Stemming in Bengali

Standard morphological parsing strategy decomposes a word into its constituent morphemes given the lexicon list, proper lexicon order and various spelling change rules. But this is not enough to compute the part of speech of a derivationally complex word or return the inflectional features of a surface level word. Existing effort in literature for Bengali is very less in number.

(Dasgupta and Khan, 2004) reported a Morphological Parser for Bengali using PC-KIMMO[1], which is widely used by linguists around the world for morphological parsing and generation. PC-KIMMO is based on Kimmo Koskenniemi's famous model of Two-level Morphology in which a word is represented as a correspondence between its lexical level form and its surface level form.

(Sarkar and Bandyopadhyay, 2009) have presented a rule-based stemming system for Bengali. At first some detail corpus study and some meaningful observation regarding Bengali orthographic stemming variation have been presented. It has been reported that noun and verb are the main two POS categories, where relatively more number of inflection are added. Finally the rules are generated by corpus study and with the help of Bengali grammar.

A.1.3 Stemming Even More Tougher for Bengali

Bengali is one of the most morphologically rich languages. Statistics shows the difficulties and various characteristics of inflections in Bengali. More than one inflection can be applied to a stem to form the surface word.

A thorough analysis of NEWS corpus used in the present work reveals that up to three inflections may be applied to a stem. Categorically, words with different POS take different number of inflections after the stem. Table A.1 presents the distinct number of inflections that can be added to a word of a specific POS as well as the number of inflections that can be added as the first, second and third inflections. The number of inflections at the n^{th} position after a stem is designated as P_n.

[1] http://www.sil.org/pckimmo/

POS	Total	P_1	P_2	P_3
Noun	34	34	6	1
Pronoun	37	34	10	2
Adjective	18	16	3	0
Adverb	8	8	3	0
Verb	129	127	4	1
Conjunction	3	3	0	0
Postposition	5	5	1	0

Table A.1: Inflection Statistics for Different POS categories in Bengali

But stemming is easier for words of closed POS types by dictionary based approaches or by other standard techniques. Stemming is a hard problem for the four open POS categories; Noun, Adjective, Adverb and Verb. But there are categorical differences in stemming among these four classes. The most general approach to solve the stemming as a problem is as follows;

1. **Lexicon**

   The list of stems and affixes, together with basic information lexicon about them (whether a stem is a Noun stem or a Verb stem, etc).

2. **Morphotactics**

   The model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes inside a word. For example, the suffixes representing the Tense, Aspect and Person information follow the Verbs rather than preceding it.

3. **Orthographic Rules**

   These spelling rules are used to model the changes that occur in a word, usually when two morphemes combine. For example, root word hAt (হাঁট) is changed into hEt (হেঁট) when added with the verb suffix to form the surface word hEtECI (হেঁটেছি)

It is very clear from the previous statistics that stemming for words with Verb POS category is the most problematic. Bengali has a vast inflectional system; the number of inflected and derivational forms of a certain lexicon is huge. For example, there are nearly (10*5) forms for certain verb words in Bengali as there are 10 tenses and 5 persons and a root verb changes its form according to tense and person. There are 20 forms of the verb root KA (খা).

Stemming is a hard problem for the morphologically rich language Bengali. But Morphological Parsing in Information Retrieval or Sentiment Analysis aspect does not demand full morphological feature structure always, rather identification of stems from several surface forms of a particular word are

required. Therefore, a Morphological stemmer based on stemming cluster technique has been developed. The details of the proposed methodologies are discussed below.

A.1.4 The Proposed Stemming Cluster based Morphological Stemmer

Two types of morphological clustering strategy have been used in the present work. The first strategy is for agglutinative suffix stripping. A manually generated suffix list is used in the present task. The list is sorted in descending order based on the length of the suffixes. The second strategy works with minimum edit distance method along with a suffix list.

A.1.4.1 Corpus-Based Acquisition of Suffix List

The suffix list used has been generated semi-automatically from the NEWS corpus used in the present work. Four separate lists have been prepared corresponding to the noun, adjective, adverb and verb POS categories. A basic clustering technique with threshold value of -3 (deletion of three characters at the end of the word) to +3 (insertion of three characters at the end of the word) has been considered to form the clusters of the words in the corpus. Every cluster centre is considered as the root form of the surface words in that cluster. A list of suffixes is generated from the surface forms of a word by subtracting the root word from the surface words. The automatically generated suffix list is then sorted to remove the duplicates and then manually checked to build up the final list. Table A.2 reports a snapshot of the semi-automatically generated suffix list prepared under various POS categories.

Type	Root	Surface Form	Suffixes
Noun	ভারত	ভারতে, ভারতের	ে, ের
Adjective	অমানব, দুর্ভাগ্য	অমানবিক, দুর্ভাগ্যবশত	িক বশত
Adverb	ভারী, দূর, দূর	ভারিক্কি, দূরীভূত	িক্কি, ীভূত
Verb	খা	খাচ্ছেন, খেয়েছিলেন	চ্ছেন, য়েছিলেন

Table A.2: Semi-Automatically Generated Suffix List

A.1.4.2 Simple Suffix Stripping

Simple suffix stripping algorithm works well for words with Noun, Adverb and Adjective classes. Each unassigned word is checked with every cluster centre after subtracting the suitable suffix from the categorical suffix list. The algorithm starts the iteration from *k* number of clusters where *k* is the total number of word forms present in a particular document.

A.1.4.3 Clustering

The stemming cluster technique analyzes the prefix and the suffix features of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as the cluster center. The term prefix/suffix is a

sequence of first/last few characters of a word, which may not be linguistically meaningful. The use of prefix/suffix information works well for highly inflected languages like the Indian languages. In case of Verbs in Bengali, root form of a word changes when suffixes are added. Hence for the Bengali Verb words simple suffix stripping does not work well. The score based stemming technique has been designed to resolve the stem for inflected Verb words. The technique uses Minimum Edit Distance method (Kukich, 1992), well known for spelling error detection, to measure the cost of identification of the class of every word. The Score based technique considers two standard operations of Minimum Edit Distance, i.e., insertion and deletion. Table A.3 (table taken from **Book:** *"Speech and Language Processing"*, Jurafsky Martin, Page-155) reports a good example of Minimum Edit Distance to compute the edit distance between the two words "*intension*" and "*execution*". The minimum edit distance between the two words as identified from the table A.3 is 8, considering the cost of insertion and deletion is 1 and the cost of substitution is 2.

n	9	10	11	10	11	12	11	10	9	**8**
o	8	9	10	9	10	11	10	9	**8**	9
i	7	8	9	8	9	10	9	**8**	9	10
t	6	7	8	7	8	9	**8**	9	10	11
n	5	6	7	6	7	**8**	9	10	11	12
e	4	5	6	**5**	**6**	7	8	9	10	11
t	3	4	**5**	6	7	8	9	10	11	12
n	2	3	**4**	5	6	7	8	8	10	11
i	1	**2**	3	4	5	6	7	8	9	10
#	**0**	1	2	3	4	5	6	7	8	9
	#	e	x	e	c	u	t	i	o	n

Table A.3: Computation of Minimum Edit Distance between "*intention*" and "*execution*".

The maximum considered range for the insertion and deletion in the present task is three characters. The idea is that the present word matches an existing cluster centre after insertion and/or deletion of maximum three characters. The present word will be assigned to the cluster that can be reached with the minimum number of insertion and/or deletion operations. This is an iterative clustering mechanism for assigning each word into a cluster. The system iterates 6 times, i.e., it starts from -3 (deletion of three characters) and ends with +3 (insertion of three characters) and finally generates a finite number of stemming clusters. A separate list of verb inflections (only 50 entries) has been maintained to validate the result of the score based technique. The standard K-means Clustering technique has been used here. K-means is a hard clustering algorithm that defines clusters by the center of mass of their members. K-means need a set of initial cluster centers in the beginning. The initial clusters are obtained from the simple suffix stripping algorithm. Then it goes through several iterations of assigning each object to the cluster whose center is closest. Examples of the Stemming cluster output for each POS category are reported in the Table A.4.

After all the words have been assigned to some cluster, a re-computation is done to identify the cluster center of each cluster as the centroid or mean of its members. The manually edited list of suffixes has been re-used here to validate the cluster members. Since the manually edited suffix list is not an exhaustive list, the words whose distance from the cluster centre is less than or equal to two characters, are kept in the cluster. Otherwise, a separate cluster is created with the word.

POS Type	Stemming Clusters
Noun	**সত্যজিৎ**,সত্যজিৎকে,সত্যজিত,সত্যজিতের
Noun	**ছবি**,ছবির,ছবিটি,ছবিতে,ছবিতেই,ছবিটিতে,ছবিটির
Noun	**রায়**,রায়ের
Post Position	**ওপর**,ওপরেও
Verb	**করত**,করতেন
Verb	**নিয়ে**,নিয়েছেন
Verb	**দেন**,দেননি
Adjective	**নির্মিত**

Table A.4: Stemming Clusters (Cluster Centre/Root Word shown in Bold)

A.1.5 Evaluation of Stemmer

Evaluation of the present Morphological Stemmer system has been done on gold standard Morphological dataset in Bengali that has been developed as part of the Government of India funded project "Indian Languages to Indian Languages Machine Translation System (IL-ILMT)[2]" The dataset consist of 1000 sentences and approximately 10K word forms. From the complete morphological output, clusters have been automatically formed by the morphological output in the gold standard dataset by looking for the same root word in the morphological feature structure. The system has reported an accuracy of 74.6%.

A.2 The Dependency Parser for Bengali

The development of a full dependency parser is indeed a separate independent research endeavor. To build the Dependency Parser we participated in the ICON 2009[3] and 2010[4] NLP TOOLS CONTEST: IL Dependency Parsing tasks. The ICON 2010 NLP TOOLS CONTEST: IL Dependency Parsing at ICON 2010

[2] http://ltrc.iiit.ac.in/ILMT/
[3] http://ltrc.iiit.ac.in/nlptools2009/
[4] http://ltrc.iiit.ac.in/nlptools2010/

datasets was provided with fine-grain and coarse-grain tagset. The details of the fine-grain and coarse-grain tagset are reported in the TableA.5.

Fine-Grain Tag	Fine-Grained Tag description	Coarse-Grain Tag description
k1	Karta (doer/agent/subject)	k1
pk1	prayojaka karta (Causer)	k1
jk1	prayojya karta (causee)	vmod
mk1	madhyastha karta (mediator-causer)	vmod
k1g	gauna karta (secondary karta)	vmod
k1s	vidheya karta (karta samanadhikarana)	k1s
k2	Karma (object/patient)	k2
k2p	Goal, Destination	k2p
k2g	gauna karma (secondary karma)	vmod
k2s	karma samanadhikarana (object complement)	k2s
k3	Karana (instrument)	k3
k4	Sampradaana (recipient)	k4
k4a	anubhava karta (Experiencer)	k4a
k5	Apaadaana (source)	k5
k5prk	prakruti apadana (‘source material' in verbs denoting change of	vmod
k7t	kaalaadhikarana (location in time)	k7
k7p	Deshadhikarana (location in space)	k7
k7	Vishayaadhikarana (location abstract)	k7
k*u	Saadrishya (similarity)	Vmod
k*s	Samanadhikarana (complement)	Vmod
r6	Shashthi (possessive)	r6
r6-k1, r6-k2	*karta* or *karma* of a conjunct verb (complex predicate)	r6-k1, r6-k2
r6v	('kA' relation between a noun and a verb)	Vmod

Fine-Grain Tag	Fine-Grained Tag description	Coarse-Grain Tag description
adv	Kriyaavisheshana ('manner adverbs' only)	Vmod
sent-adv	Sentential Adverbs	vmod
rd	Prati (direction)	Vmod
rh	Hetu (cause-effect)	Rh
rt	Taadarthya (purpose)	Rt
ras-k*	upapada__ sahakaarakatwa (associative)	Vmod
ras-neg	Negation in Associatives	Vmod
rs	relation samanadhikaran (noun elaboration)	Rs
rsp	relation for duratives	Nmod
rad	Address words	Vmod
nmod__relc, jjmod__relc, rbmod__relc	Relative clauses, jo-vo constructions	Relc
nmod__*inv		Nmod
nmod	Noun modifier (including participles)	Nmod
vmod	Verb modifier	Vmod
jjmod	Modifiers of the adjectives	Jjmod
rbmod	Modiiers of adverbs	Rbmod
pof	Part of relation	Pof
ccof	Conjunct of relation	Ccof
fragof	Fragment of	Fragof
enm	Enumerator	Vmod
nmod__adj	adjectival modifications	nmod__adj
lwg__psp	noun and post-position/suffix modification	lwg__psp
lwg__neg	NEG and verb/noun modification	lwg__neg
lwg__vaux	Auxiliary verb modification	lwg__vaux
lwg__rp	particle modification	lwg__rp
lwg_cont	lwg continuation relation	lwg_cont
lwg__*	Other modifications in lwg	lwg__rest
jjmod__intf	intensifier adjectival modifications.	jjmod__intf
pof__redup	reduplication	pof__redup
pof__cn	compound noun	pof__cn
pof__cv	compound verb	pof__cv
rsym	Punctuations and symbols	Rsym
mod	Modifier	mod

Table A.5: Dependency Tag Set

Due to syntactic richness of Bengali, hybrid architecture has been proposed in the present task. A statistical data driven parsing system (Maltparser[5]) has been used followed by a rule-based post-processing technique. The system has been trained on the ICON NLP TOOLS CONTEST: IL Dependency Parsing datasets. The final system (trained on ICON 2010 NLP TOOLS CONTEST Dataset) demonstrated an accuracy of unlabeled attachment score (UAS): 83.87%, labeled attachment score (LAS): 64.31% and labeled accuracy score (LS): 69.3% respectively over the fine-grained tagset.

Bengali is characterized by a rich system of inflections (VIBHAKTI), derivation and compound formation (Saha et al., 2004; Chakroborty, 2003) and ***karakas***, which is why the Natural Language Processing tasks for Bengali are very challenging. These language specific peculiarities play important roles in the parsing of natural language sentences. The development of a computational grammar for a natural language, also known as Grammar development or Grammar Engineering, can be a complex task.

Previous research efforts have proposed two different approaches in the context of parsing of natural language sentences. These techniques are known as grammar driven parsing and data driven parsing. Most of the previous grammar driven parsing research attempts were for detection and formation of the proper rule set to identify the characteristics of inter-chunk relations.

The development of a proper set of parsing rules will always remain inadequate. Most of the modern grammar-driven dependency parsers (Karlsson et. al.; 1995, Bharati et. al., 2008) parse by eliminating the parses which do not satisfy the given set of constraints. Thus, the data driven parsing system (Maltparser[6] ver.1.3.1) has been considered as the baseline in the present task. The data driven parser requires a large set of manually annotated corpus. But the available dataset is not large enough in size. Therefore, a hybrid technique has been proposed that filters the output of the baseline system by a rule-based post-processing system.

On the other hand, in our previous endeavor (Ghosh et. al, 2009) for ICON 2009 NLP TOOLS CONTEST: IL Dependency Parsing task, experiments have been carried out with a statistical Conditional Random Field (CRF) based model followed by rule-based post-processing techniques. The system has demonstrated an unlabeled attachment score (UAS) of 74.09%, labeled attachment score (LAS) of 53.90% and labeled accuracy score (LS) of 61.71% respectively. The evaluation results of our Dependency parser that participated in the ICON 2010 NLP TOOLS CONTEST: IL Dependency Parsing task outperforms the previous attempt. The details of the ICON 2010 NLP TOOLS CONTEST task have been mentioned in the subsequent sections.

The standard dependency evaluation metrics like Unlabeled Attachment Score (UAS), Labeled Accuracy score (LS), and Labeled Attachment Score (LAS) have been used to evaluate the dependency parser's accuracy (Nivre et. al., 2007a). UAS is the percentage of words in the sentences across the entire test

[5] http://maltparser.org/
[6] http://maltparser.org/download.html

data that have correct parents. LS is the percentage of words with correct dependency labels, while LAS is the percentage of words with correct parent and correct dependency labels.

A.2.1 Dataset

The dataset for the ICON 2010 NLP TOOLS CONTEST: IL Dependency Parsing task at ICON 2010 was provided with fine-grain and coarse-grain tagsets. The corpus statistics is reported in the Table A.6. A few detailed statistics about the distribution of sentence types in the corpus is reported in Table A.7.

	Sentences	Tokens	Number of Tokens Per Sentence
Training	960	7269	7.57
Development	150	812	5.41
Testing	150	962	6.4

Table A.6: Corpus Statistics of ICON 2010 NLP Tools Contest

Corpus	simple	compound	Complex
Training	223	188	589
Development	31	11	108
Testing	26	7	117

Table A.7: ICON 2010 NLP Tools Contest Corpus Statistics on Sentence Types

A.2.2 Using the Maltparser

The Maltparser uses a classifier based Shift/Reduce parsing methodology. It uses arc-eager, arc-standard, covington projective and convington non-projective algorithms for parsing (Nivre, 2006). History-based feature models are used for predicting the next parser action (Black et. al., 1992). Support Vector Machine (SVM) is used for mapping histories to parser actions (Kudo and Matsumoto, 2002). It uses graph transformation to handle non-projective trees (Nivre and Nilsson, 2005).

Maltparser accepts CoNLL format[7] as the input. Six morphological features namely lexicon, morphological category, gender, number, person, vibhakti or Tense-Aspect-modality (TAM) markers of the node are considered in the present set of experiments. After experimentation with different sets of feature combinations, the vibhakti, TAM and the morphological category produce better results as these features contain most crucial information to identify dependency relations for Indian languages and especially for Bengali.

The Bengali dataset consists of 7% non-projective (Nivre, 2009) sentences. Among the four parsing algorithms provided with Maltparser, it has been found that nivreeager (Nivre, 2009) works best for the Bengali corpus with the fine-grained tagset. Analyzing the parser output with default setting, it has been

[7] http://ilk.uvt.nl/conll/

found that the parsing of complex and compound sentences generate most of the errors. The average number of tokens per sentence in the corpus is calculated as 6. Thus the maximum sentence length was set to 6.

The Maltparser uses the Support Vector Machine (SVM) learning method with polynomial kernel to map the feature vector representation of a parser configuration. While tuning the learning method parameter, the cost parameter of the Maltparser was changed from the default value of 1 to .65, which controls the tradeoff between minimizing training error and maximizing margin. The corpus statistics revealed that the average number of tokens per sentence is 6 and the number of attributes per node or chunk is 9. Thus, experiments were also carried out with the Liblinear (Nivre, 2009) classifier. Due to the small size of the dataset, many dependency relations are sparsely distributed, which leads to low LAS value. The comparative study of the accuracies of different Maltparser configurations is shown in the Table A.8.

Algorithm	UAS[8]	LAS[9]	LS[10]
Nivreeager+Liblinear	**81.64%**	**54.58%**	**50.62%**
Convington non-projective+LIBSVM	78.22%	51.02%	48.43%
Convington non-projective+ Liblinear	79.33%	52.47%	50.52%

Table A.8: Comparison of Maltparser Output with Different Settings

The confusion matrix on the development set for some important dependency relations is shown in the Table A.9.

	k1	k2	k7p	k7t	pof	vmod
k1	0	29	3	1	6	1
k2	7	0	0	0	6	9
k7p	12	6	0	3	1	2
k7t	3	2	0	0	1	4
Pof	2	3	0	0	0	0
Vmod	4	7	1	2	0	0

Table A.9: Confusion Matrix on Development Set

[8] UAS – Unlabeled Attachment Score
[9] LAS – Labeled Attachment Score
[10] LS - Labeled Score

A.2.3 Post-Processing

With the help of confusion matrix, a set of post-processing rules has been devised depending on the nature of errors. Vibhakti plays a crucial role in the identification of dependency relations. As the vibhakti information is missing in some cases in the corpus, a suffix analyzer (with a manually augment list) is applied to the word to identify the vibhakti / inflection. Some of these post-processing rules for some important dependency relation tags are described below.

- **r6:** The r6 dependency relation tag denotes ***genitive relation***. It takes 'র', 'ের' and 'দের' as the genetive markers. For the marker 'র', it can appear at the end of many words, e.g. 'আবিষ্কার'. A dictionary based approach has been used to exclude such words. When chunks with these genitive markers have any indirect relation with the main verb, then it is marked with r6 dependency relation tag and the following NP chunks are marked as the related chunks.

- **k7t:** The k7t dependency relation tag denotes ***time temporal***. Generally, such chunks take the suffix 'কালে'. A list of time temporal words has been manually developed from the training corpus. The list is used to identify the time temporal chunks. Such chunks are marked with k7t dependency relation tags.

- **k7p:** The k7p dependency relation tag denotes ***space temporal.*** Generally, such chunks take the suffix 'পুর'. A list of space temporal chunks has been manually developed from the training corpus. The list is used to identify the space temporal chunks. Such chunks are marked with k7p dependency relation tags.

After an in depth study of the errors made by Maltparser, a rule based system has been developed with the help of linguistic knowledge. Depending on the specific attributes of a chunk like vibhakti/case markers and/or word information, the rule based system derives the dependency relations of the chunk. For each dependency relation tag, syntactic cues are derived to identify the dependency relations depending on specific linguistic features. Some of these syntactic cues are,

1. A NP chunk with null vibhakti and NNP or PRP POS tag will be marked with k1 relation with the nearest verb chunk.

2. A chunk head with "র" vibhakti will be marked with 'r6' relation with the next noun chunk.

3. A NP chunk with null vibhakti and NN POS tag will be marked with k2 relation with the nearest verb chunk.

4. In co-ordinate type sentences, the verb chunk will be marked with 'ccof' relation with the nearest CCP chunk. If the CCP chunk is surrounded by two NP chunks then the NP chunks will be marked with 'ccof' relation with the CCP chunk.

5. Sub-ordinate sentences are identified based on the presence of keywords like 'যে' etc. In sub-ordinate type sentences, the verb chunk of the sub-ordinate clause will be marked with "nmod__relc" relation with that chunk of the main clause, which the sub-ordinate clause is modifying.
6. A NP chunk with "0_থেকে" vibhakti will be marked with k5 relation with the nearest verb chunk.
7. A NP chunk with "0_প্রতি" vibhakti will be marked with 'rd' relation with the nearest verb chunk.
8. After carefully analyzing the training corpus, certain vibhaktis or rather post-position markers with semantic meanings have been found like '0_পক্ষে', '0_হিসাবে' etc. that can be treated as cues to mark the 'vmod' relation.
9. Verb like 'কর' or 'হ' often takes another argument to form compound verbs. The argument is marked with part-of relation (pof). The preceding noun or verb chunk, if it has no suffix, is marked with 'pof' relation.
10. If a NP chunk is marked with "কে" vibhakti, 'k2' relation will be identified.
11. Noun chunks with root words like "আমি" with "NN" POSt ag or "তুমি" with "PRP" POS tag will be marked with 'k1' relation.
12. If the root word is 'যে' and the word is 'যা', 'তা', then the chunk will be marked with 'k2' relation.

The ambiguity comes when for a certain vibhakti, multiple possible relations are identified. For example, chunks with null vibhakti to the chunk headword can have two possible dependency relations, 'k1' and 'k2'. The ambiguity is resolved using the POS tag. If the POS tag is 'NNP' then the dependency relation will be 'k1' and if the POS tag is 'NN' then the dependency relation will be 'k2'. If ambiguity is not resolved with this rule then the position of the chunk in the sentence is considered. If there are two chunks with null vibhakti, the chunk distant from the verb chunk will be marked with 'k1' relation and the nearer one chunks will be marked with 'k2' relations.

After studying the co-occurrence pattern of 'k1' and 'k2' relations in a sentence, it has been observed that the single occurrence of noun chunk with the null vibhakti is marked with the 'k1' relation.

The output of the Maltparser and the output of the rule based parsing system are compared. The rule based system is given the higher priority as it is based on syntactic-semantic cues. If there is any mismatch between the results from the two systems and if the rule based system has generated an output then the output of the rule based system is considered.

A.2.4 Performance of the Dependency Parser

The Maltparser has been trained with the training dataset with fine-grained tagset only. The Maltparser with nivreeager parsing algorithm, yielded unlabelled attachment score of 81.64%, Labeled attachment

score of 54.58% and Labeled score of 50.62% on the development set. After the application of the suffix analyzer, 3% improvement of UAS and 9% improvement of LAS scores have been achieved. The LS score jumps to 69%. The rule based system yields UAS, LAS, LS scores of 84.02%, 66.63% and 70.82% respectively on the development set.

In the final evaluation on the test dataset as part of the ICON 2010 tool contest, the system has demonstrated UAS (Unlabelled Accuracy Score) score of 83.87%, LAS (Labeled Accuracy Score) score of 64.31% and LS (Labeled Score) score of 69.3% respectively.

Publications

1. **Amitava Das** and Sivaji Bandyopadhyay. 2010. ***Morphological Stemming Cluster Identification for Bangla*.** *In Knowledge Sharing Event-1: Task 3: Morphological Analyzers and Generators*, January 24-25, 2010, Mysore, India.
   http://www.amitavadas.com/Pub/Morph_Bengali.pdf

2. Aniruddha Ghosh, **Amitava Das**, Pinaki Bhaskar and Sivaji Bandyopadhyay. 2010. ***Bengali Parsing System at ICON NLP Tool Contest 2010***. *In the NLP Tool Contest*: *Dependency Parsing at International Conference on Natural Language Processing (ICON 2010),* Pages 20-24, Kharagpur, India.
   http://researchweb.iiit.ac.in/~prashanth/papers/husain-mannem-ambati-gadde-icon10.pdf
   http://www.sivajibandyopadhyay.com/pinaki/papers/JU%20parsing%20system-2010.pdf
3. Aniruddha Ghosh, **Amitava Das**, Pinaki Bhaskar and Sivaji Bandyopadhyay. 2009. ***Dependency Parser for Bengali: the JU System at ICON 2009***. *In the NLP Tool Contest*: *Dependency Parsing at International Conference on Natural Language Processing (ICON 2009)*, Pages 7-11, Hyderabad, India.
   http://ltrc.iiit.ac.in/nlptools2009/CR/all-papers-toolscontest.pdf
   http://ltrc.iiit.ac.in/nlptools2009/CR/Parser_Camera_Ready_JU.pdf

Bibliography

Abbasi, Ahmed; Chen, Hsinchun and Salem, Arab. 2008. **Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums**. In the ACM Transactions on Information Systems, 26(3), No. 12.

http://ai.arizona.edu/intranet/papers/ahmedabbasi_sentimenttois.pdf

Agrawal, R.; Bayardo, R.; Gruhl, D. and Papadimitriou S. 2002. **Vinci: A Service-Oriented Architecture for Rapid Development of Web Applications**. In the Proceedings of the WWW Conference, Pages 355–365, Hawaii, USA.

www.almaden.ibm.com/cs/people/dgruhl/vinci.pdf

Akkaya, Cem; Wiebe, Janyce; Conrad, Alexander and Mihalcea, Rada. 2011. **Improving the Impact of Subjectivity Word Sense Disambiguation on Contextual Opinion Analysis**. In the Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL 2011), Pages 87-96, Portland, Oregon, USA.

http://www.aclweb.org/anthology/W11-0311.pdf

Andreevskaia, Alina and Bergler, Sabine. 2007. **Clac and Clac-NB: Knowledge-Based and Corpus-Based Approaches to Sentiment Tagging**. In the Proceeding of the 4th SemEval-2007, ACL 2007, Pages 117–120, Prague, Czech Republic.

http://portal.acm.org/citation.cfm?id=1621474.1621496

Aue, Anthony and Gamon, Michael. 2005. **Customizing Sentiment Classifiers to New Domains: a Case Study**. In the Proceeding of the International Conference on Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria.

http://research.microsoft.com/~anthaue/new_domain_sentiment.pdf

Auria, L. and Moro, Rouslan A. 2008. **Support Vector Machines (SVM) as a Technique for Solvency Analysis**. ISSN print edition 1433-0210, Berlin.

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1424949

Baccianella S., Andrea E. and Sebastiani, F. 2010. **SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining**. In the Proceedings of the 7th conference on International Language Resources and Evaluation (LREC-10), European Language Resources Association (ELRA), Pages 2200-2204, Valletta, Malta.

http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf

Bal, Krishna Bal and Saint-Dizier, Patrick. 2010. **Towards Building Annotated Resources for Analyzing Opinions and Argumentation in News Editorials**. In LREC, Malta.

www.lrec-conf.org/proceedings/lrec2010/pdf/677_Paper.pdf

Beineke, P.; Hastie, T. and Vaithyanathan, S. 2004. **The Sentimental Factor: Improving Review Classification Via Human-Provided Information**. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 263.

http://acl.ldc.upenn.edu/P/P04/P04-1034.pdf

Bethard, Steven; Yu Hong; Thornton, Ashley; Hatzivassiloglou, Vasileios and Dan, Jurafsky. 2006. **Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues**. In the Computing Attitude and Affect in Text: Theory and Applications, Pages 125-141.

http://www.stanford.edu/~jurafsky/bethard_05_opinion_propositions.pdf

Bibliography

Bharati, A.; Husain, S.; Ambati, B.; Jain S.; Sharma, D.M. and Sangal, R. 2008. **Two semantic features make all the difference in Parsing accuracy**. In the Proceedings of the 6th International Conference on Natural Language Processing (ICON-08), CDAC Pune, India.

http://researchweb.iiit.ac.in/~ambati/papers/Icon08-parser-semfeat-camera_ready.pdf

Black E., Jelinek F., Lafferty J. D., Magerman D.M., Mercer R. L. and Roukos S. 1992. **Towards History-Based Grammars: Using Richer Models For Probabilistic Parsing**. In Proceeding of the 5th DARPA Speech and Natural Language Workshop, Pages 31–37.

http://acl.ldc.upenn.edu/P/P93/P93-1005.pdf

Bloom Kenneth, Stein Sterling and Argamon Shlomo. 2007. **Appraisal Extraction for News Opinion Analysis at NTCIR-6**. In the Proceedings of NTCIR-6 Workshop Meeting, Tokyo, Japan.

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/45.pdf

Cambria Erik, Hussain Amir and Eckl Chris. 2011. **Taking Refuge in Your Personal Sentic Corner**. 2011. In the Proceeding of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), (IJCNLP 2011), Pages 35-43, Chiang Mai, Thailand.

www.aclweb.org/anthology/W11-3706

Carenini Giuseppe; Ng, Raymond and Pauls, Adam. 2006. **Multi-document Summarization of Evaluative Text**. In the Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), Pages 305–312, Trento, Italy.

http://acl.ldc.upenn.edu/E/E06/E06-1039.pdf

Carenini, Giuseppe; Ng, Raymond T. and Pauls, Adam. 2006. **Interactive Multimedia Summaries of Evaluative Text**. In the Proceedings of the 11th International Conference on Intelligent User Interfaces, Pages 124–131, Sydney, Australia.

www.eecs.berkeley.edu/~adpauls/PAPERS/iui06.pdf

Chakroborty, B. 2003. **Uchchotoro Bangla Byakaron**. Published by AkshayMalancha.

Chesley, Paula; Vincent, Bruce; Xu, Li, and Srihari, Rohini. 2006. **Using Verbs and Adjectives to Automatically Classify Blog Sentiment**. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), Pages 27–29, Stanford, USA.

www.cs.pitt.edu/~wiebe/courses/CS3730/Fall08/chesleyetal2005.pdf

Choi Yejin, Cardie Claire, Riloff Ellen and Patwardhan Siddharth. 2005. **Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns**. In the Proceedings of the HLT/EMNLP 2005, Pages 355-362, Vancouver, British Columbia, Canada.

www.cs.utah.edu/~riloff/pdfs/emnlp05.pdf

Choi, Yoonjung and Kim, Youngho and Myaeng, Sung-Hyon. 2009. **Domain-specific Sentiment Analysis using Contextual Feature Generation**. In the Proceeding of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, 2009, Hong Kong, China.

http://ir.kaist.ac.kr/anthology/2009.11-Choi.pdf

Clynes, M. 1997. **Sentics: The Touch of the Emotions**. Doubleday, New York.

http://senticcycles.org/home/sentics/articles/sentics.pdf

Das, Sanjiv R. and Chen, Mike Y. 2007. **Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web**. In Management Science, 53(9), Pages 1375–1388.

http://algo.scu.edu/~sanjivdas/chat_final.pdf

Dasgupta, Sajib and Khan, Mumit. 2004. **Feature Unification for Morphological Parsing in Bangla**. *In the Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT),* Bangladesh.

http://www.utdallas.edu/~sxd052000/feature2.pdf

Dasgupta, Sajib and Ng, Vincent. 2009. **Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification**. In the conference of the EMNLP, Singapore.

http://www.utdallas.edu/~sxd052000/EMNLP-09-Sajib.pdf

Dave, Kushal and Lawrence, Steve and Pennock, David M. 2003. **Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews**. In the Proceedings of the 12th International Conference on World Wide Web (WWW 2003), Pages 519–528, Budapest, Hungary.

http://ws.csie.ncku.edu.tw/login/upload/2006/paper/Opinion%20Extraction%20and%20Semantic%20Classification%20of%20Product%20Reviews.pdf

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. 1990. **Indexing By Latent Semantic Analysis**. In Journal of the American Society for Information Science (JASIS), 41 (6), Pages 391-407.

http://www.cs.bham.ac.uk/~pxt/IDA/lsa_ind.pdf

Denecke, Kerstin. 2009. **Are SentiWordNet Scores Suited For Multi-Domain Sentiment Classification?** In the Proceeding of the 4th International Conference on Digital Information Management (ICDIM 2009), Pages 33-38, Ann Arbor, USA.

http://www.l3s.de/web/upload/documents/1/ICDIM_final.pdf

Denhire G. and Lemaire B. 2004. **A Computational Model of Children's Semantic Memory**. In the Proceedings of the 26th Annual Meeting of the Cognitive Science Society, Pages 297–302, Chicago, USA.

http://webu2.upmf-grenoble.fr/LPNC/resources/benoit_lemaire/cogsci04_1.pdf

Dowty, David R. 1991. **Thematic Proto-Roles and Argument Selection**. Language, 67(3):547–619.

http://linguistics.berkeley.edu/~syntax-circle/syntax-group/dowty91.pdf

Efron, M. 2004. **Cultural Orientations: Classifying Subjective Documents by Cocitation Analysis**. *In Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music, and Design*, Pages 41-48.

http://www.aaai.org/Papers/Symposia/Fall/2004/FS-04-07/FS04-07-007.pdf

Ekbal A. and Bandyopadhyay S. 2010. **Voted NER System using Appropriate Unlabeled Data**. In the Lingvisticae Investigationes Journal, John Benjamins Publishing Company.

http://www.mt-archive.info/NEWS-2009-Ekbal.pdf

Ekbal, A. and Bandyopadhyay, S. 2008. **A Web-based Bengali News Corpus for Named Entity Recognition**. In Language Resources and Evaluation (LRE) Journal, Springer, Vol. 42(2), PP.173-182.

http://www.springerlink.com/content/q311u18875wkll65/

Fei, Z., Liu, J., and Wu, G. 2004. **Sentiment Classification using Phrase Patterns**. In Proceedings of the 4th IEEE International Conference on Computer Information Technology, Pages 1147-1152.

http://dl.acm.org/citation.cfm?id=1025457

Ferret, Olivier and Zock, Michael. 2006. **Enhancing Electronic Dictionaries with an Index Based on Associations**. In the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006), Pages 281-288, Sydney, Australia.

http://dl.acm.org/citation.cfm?id=1220211

Bibliography

Fillmore Charles, Johnson Christopher and Petruck Miriam. 2003. **Background to FrameNet**. International Journal of Lexicography, 16, Pages 235–250.

http://ijl.oxfordjournals.org/content/16/3/235.abstract

Fillmore, Charles. 1968. **The Case for Case**. In Emmon Bach and Robert T. Harms (eds.) Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston. Pages 1-88.

http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED019631

Fukuhara, Tomohiro; Nakagawa, Hiroshi and Nishida, Toyoaki. 2007. **Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events**. In the Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

http://www.race.u-tokyo.ac.jp/~fukuhara/Research/paper/07/fukuhara-icwsm07-camera.pdf

Gamon, Michael and Aue, Anthony. 2005. **Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms**. In the Proceedings of the Workshop on Feature Engineering for Machine Learning in Natural Language Processing, ACL-05, Pages 57–64, Ann Arbor, USA.

http://research.microsoft.com/pubs/65462/sentiment_feats_camera.pdf

Gamon, Michael, Aue, Anthony, Corston-Oliver, Simon and Ringger, Eric. 2005. **Pulse: Mining Customer Opinions from Free Text**. In the Proceedings of the International Symposium on Intelligent Data Analysis (IDA), Number 3646 in Lecture Notes in Computer Science, Pages 121–132.

http://research.microsoft.com/apps/pubs/default.aspx?id=65463

Gildea Daniel, Jurafsky Daniel. 2002. **Automatic Labeling of Semantic Roles**. In Computational Linguist. Vol. 28, No. 3. (1 September 2002), Pages 245-288.

www.cs.rochester.edu/~gildea/gildea-cl02.pdf

Goldberg, David E. 1989. **Genetic Algorithms in Search, Optimization and Machine Learning**. Addison-Wesley Longman Publishing Co., Boston, MA, USA.

http://dl.acm.org/citation.cfm?id=534133

Grassi, Marco. 2009. **Developing HEO Human Emotions Ontology**. In Proceedings of the 2009 Joint International Conference on Biometric ID management and Multimodal Communication, volume 5707 of Lecture Notes in Computer Science, Pages 244–251. Springer, Berlin.

http://www.springerlink.com/content/n3p0jk10514vt120/

Grefenstette, G., Qu, Y., Shanahan, J. G. and Evans, D. A. 2004. **Coupling Niche Browsers and Affect Analysis For An Opinion Mining Application.** In Proceedings of the 12th International Conference Recherche d'Information Assistee par Ordinateur, Pages 186-194.

http://www.cs.pitt.edu/%7Ewiebe/courses/CS3730/Fall08/grefenstetteetalriao-04.pdf

Gregory, Michelle L.; Chinchor, Nancy; Whitney, Paul; Carter, Richard; Hetzler, Elizabeth and Turner, Alan. 2006. **User-Directed Sentiment Analysis: Visualizing the Affective Content of Documents**. In the Proceedings of the Workshop on Sentiment and Subjectivity in Text, ACL 2006, Pages 23–30, Sydney, Australia.

http://dl.acm.org/citation.cfm?id=1654645

Gruhl, D. and Chavet, L. and Gibson, D. and Meyer, J. and Pattanayak, P. and Tomkins, A. and Zien, J. 2004. **How to Build a Webfountain: an Architecture For Very Large-Scale Text Analytics**. In IBM Systems Journal, 43(1), Pages 64–77.

www.ttivanguard.com/sfreconn/webfountain.pdf

Haghighi Aria, Toutanova Kristina and Manning Christopher. 2005. **A Joint Model for Semantic Role Labeling**. In CoNLL-2005 Shared Task.

http://nlp.stanford.edu/~manning/papers/conll2005new.ps

Harris, Z. 1954. **Distributional Structure**. Word, 10 (23), Pages 146-162.

http://books.google.co.in/books?id=2x8zfDFYivAC&pg=PA3&lpg=PA3&dq=distributional+structure+harris&source=bl&ots=9hj6GplcG9&sig=A2X_KHlc1HnT5fEE7w9xioHT5TA&hl=en&ei=AI3KTtWOCoaGrAfPoLm7Dg&sa=X&oi=book_result&ct=result&resnum=9&ved=0CFoQ6AEwCA#v=onepage&q&f=false

Haruechaiyasak, Choochart; Kongthon, Alisa; Palingoon, Pornpimon and Sangkeettrakarn, Chatchawal. 2010. **Constructing Thai Opinion Mining Resource: A Case Study on Hotel Reviews**. In the Proceedings of the 8th Workshop on Asian Language Resources, Pages 64–71, Beijing, China.

http://www.aclweb.org/anthology/W10-3209

Hatzivassiloglou, V. and McKeown, K. 1997. **Predicting the Semantic Orientation of Adjectives**. In the Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, Pages 174–181, Madrid, Spain.

http://acl.ldc.upenn.edu/P/P97/P97-1023.pdf

Hatzivassiloglou, Vasileios and Wiebe, Janyce. 2000. **Effects of Adjective Orientation and Gradability on Sentence Subjectivity**. In the Proceedings of the International Conference on Computational Linguistics (COLING 2000), Pages 299-305, Saarbrücken, Germany.

www.aclweb.org/anthology/C/C00/C00-1044.pdf

Havasi, C., Speer, R., Alonso, J. 2007. **ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge**. In the Proceeding of the Recent Advances in Natural Languages Processing (RANLP 2007).

www.media.mit.edu/~jalonso/cnet3.pdf

He Yulan, Alani Harith and Zhou Deyu. 2010. **Exploring English Lexicon Knowledge for Chinese Sentiment Analysis**. In the Proceeding of the Joint Conference on Chinese Language Processing (CIPS-SIGHAN), Beijing, China.

www.aclweb.org/anthology/W/W10/W10-4116.pdf

He, Ben; Macdonald, Craig; Ounis, Iadh; Peng, Jie and Santos, Rodrygo L.T. 2008. **University of Glasgow at TREC 2008: Experiments in Blog, Enterprise, and Relevance Feedback Tracks with Terrier**. In the Proceedings of the 7th Text Retrieval Conference (TREC 2008).

http://trec.nist.gov/pubs/trec17/papers/uglasgow.blog.ent.rf.rev.pdf

Hetzler, Elizabeth and Turner, Alan. 2004. **Analysis Experiences Using Information Visualization.** In IEEE Computer Graphics and Applications, 24(5), Pages 22-26.

http://dl.acm.org/citation.cfm?id=1024888

Holland, John H. 1975. **Adaptation in Natural and Artificial Systems**. MIT Press, Cambridge, MA, USA.

http://dl.acm.org/citation.cfm?id=129194 ,
http://www.owlnet.rice.edu/~psyc351/Lectures/Jewell351EvolutionLecture2002.pdf

Hu, Minqing and Liu, Bing. 2004. **Mining and Summarizing Customer Reviews**. In the Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 168–177, Seattle, WA, USA.

www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf

Bibliography

Hugo Liu, Henry Lieberman, Ted Selker. 2003. **A Model of Textual Affect Sensing using Real-World Knowledge**. In the Proceedings of the International Conference on Intelligent User Interfaces (IUI 2003), Pages 125-132, Miami, Florida.

http://www.media.mit.edu/~hugo/publications/papers/IUI2003-affectsensing.pdf

Jia, Lifeng; Yu, Clement; Zhang, Wei. 2008. **UIC at TREC 2008 Blog Track**. In the Proceedings of the 7th Text Retrieval Conference (TREC 2008).

http://trec.nist.gov/pubs/trec17/papers/uillinois-chicago.blog.pdf

Jurafsky, Daniel and Martin, James. 1999. Speech and Language Processing.

Kanayama, H., Nasukawa, T., and Watanabe, H. 2004. **Deeper Sentiment Analysis Using Machine Translation Technology**. *In Proceedings of the 20th International Conference on Computational Linguistics*, Pages 494- 500.

http://acl.ldc.upenn.edu/C/C04/C04-1071.pdf

Karlsson, F., A. Voutilainen, J. Heikkilä and A. Anttila (eds). **Constraint Grammar: A Language Independent System for Parsing Unrestricted Text**. Mouton de Gruyter. 1995.

http://books.google.co.in/books?id=70IvVPlH63cC&pg=PA1&lpg=PA1&dq=Constraint+Grammar:+A+Language+Independent+System+for+Parsing+Unrestricted+Text&source=bl&ots=mB6pymoLL9&sig=IiYBuJRhZr1zxh86NRHHN-e-XPw&hl=en&ei=rJvKTqj4M5HJrQexz-zzAw&sa=X&oi=book_result&ct=result&resnum=8&ved=0CE0Q6AEwBw#v=onepage&q=Constraint%20Grammar%3A%20A%20Language%20Independent%20System%20for%20Parsing%20Unrestricted%20Text&f=false

Kawai Yukiko, Kumamoto Tadahiko and Tanaka Katsumi. 2007. **Fair News Reader: Recommending News Articles with Different Sentiments Based on User Preference**. In the Knowledge-Based Intelligent Information & Engineering Systems (KES), Pages 612-622.

www.springerlink.com/index/ynhr25513w122214.pdf

Kenji, Tateishi; Yoshihide, Ishiguro and Toshikazu, Fukushima. 2001. **Opinion information retrieval from the Internet**. In the Information Processing Society of Japan (IPSJ) SIG Notes, 2001(69(20010716)):75–82, 2001.

http://sciencelinks.jp/j-east/article/200121/000020012101A0840759.php

Kim Soo-Min and Hovy Eduard. 2004. **Determining the Sentiment of Opinions**. In the Proceedings of the COLING (2004), Pages 1367-1373.

http://www.isi.edu/natural-language/people/hovy/papers/04Coling-opinion-valences.pdf

Kiparsky, Paul and J. F. Staal. 1969. **Syntactic and semantic relations in Panini**. Foundations of Language 5, Pages 83-117.

www.jstor.org/stable/25000364

Kipper Karin, Korhonen Anna, Ryant Neville and Palmer Martha. 2006. **Extending VerbNet with Novel Verb Classes.** In the Proceeding of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy.

www.cl.cam.ac.uk/~alk23/lrec06.pdf

Kleinberg, Jon and Tardos, Eva.2006. **Algorithm Design**. Addison Wesley.

http://www.aw-bc.com/info/kleinberg/

Kobayashi Nozomi, Ryu Iida, Kentaro Inui and Yuji Matsumoto. 2006. **Opinion Mining as Extraction of Attribute-Value Relations**. In Lecture Notes in Artificial Intelligence, Vol. 4012, Pages 470-481, Springer-Verlag.

http://www.springerlink.com/content/p702302020k57302/

Kraft, D.H. Petry, F.E. Buckles, B.P. Sadasivan, T. 1994. **The Use of Genetic Programming to Build Queries for Information Retrieval.** In Proceedings of the Evolutionary Computation, 1st IEEE Conference on World Congress on Computational Intelligence, Pages 468-473, Orlando, Florida.

http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=349905

Ku Lun-Wei, Liang Yu-Ting and Chen Hsin-Hsi. 2006. **Opinion Extraction, Summarization and Tracking in News and Blog Corpora**. In the Proceedings ofAAAICAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs.

http://nlg18.csie.ntu.edu.tw:8080/lwku/SS0603KuLW.pdf

Ku, Lun-Wei and Lee, Li-Ying and Wu, Tung-Ho and Chen, Hsin-Hsi. 2005. **Major Topic Detection and Its Application to Opinion Summarization**. In the Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pages 627-628, Salvador, Brazil.

http://dl.acm.org/citation.cfm?id=1076161

Kudo, T. and Matsumoto, Y. 2002. **Japanese Dependency Analysis Using Cascaded Chunking**. In CoNLL-2002. Pages 63–69.

http://acl.ldc.upenn.edu/W/W02/W02-2016.pdf

Kukich, Karen. 1992. **Techniques for Automatically Correcting Words in Text**. In the ACM Computing Surveys, Pages 377-439.

http://dc-pubs.dbs.uni-leipzig.de/files/Kukich1992Techniqueforautomatically.pdf

Landauer, T. K., & Dumais, S. T. 1997. **A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction and Representation of Knowledge**. Psychological Review, 104 (2). Pages 211-240.

http://www.stat.cmu.edu/~cshalizi/350/2008/readings/Landauer-Dumais.pdf

Landauer, T. K., McNamara, D. S., Dennis, S. and Kintsch, W. 2007. **Handbook of Latent Semantic Analysis**. Lawrence Erlbaum, Mahwah, NJ.

http://www.scholarpedia.org/article/Latent_semantic_analysis,

http://www.amazon.com/Handbook-Semantic-University-Institute-Cognitive/dp/0805854185

Lee, Yeha; Na, Seung-Hoon; Kim, Jungi; Nam, Sang-Hyob; Jung, Hun-young and Lee, Jong-Hyeok. 2008. **KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval**. In the Proceedings of the 7th Text Retrieval Conference (TREC 2008).

http://trec.nist.gov/pubs/trec17/papers/pohang.blog.rev.pdf

Li, Hang and Yamanishi, Kenji. 2001. **Mining from Open Answers in Questionnaire Data**. In the Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Pages 443–449.

http://research.microsoft.com/en-us/people/hangli/yamanishi-li-ieee-is2002.pdf

Liu, Bing. 2010. **Sentiment Analysis: A Multi-Faceted Problem**. In the IEEE Intelligent Systems, Volume 25(3), Pages 76-80.

http://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf

Bibliography

Liu, Bing; Hu Minqing and Cheng, Junsheng. 2005. **Opinion Observer: Analyzing and Comparing Opinions on the Web**. In the Proceeding of the International World Wide Web (WWW 2005), Pages 342-351, Chiba, Japan.

http://www2005.org/cdrom/docs/p342.pdf

Liu, H.; Lieberman, H.; Selker, T. 2003. **A Model of Textual Affect Sensing using Real-World Knowledge**. In the Proceedings of the 8th international conference on Intelligent user interfaces (IUI 2003), Pages 125-132, Miami, Florida, USA.

http://web.media.mit.edu/~hugo/publications/papers/IUI2003-affectsensing.pdf

Lloyd, Levon; Kechagias, Dimitrios and Skiena, Steven. 2005. **Lydia: A System for Large-Scale News Analysis**. In the Proceedings of String Processing and Information Retrieval (SPIRE), Number 3772 in Lecture Notes in Computer Science, Pages 161–166.

www.cs.sunysb.edu/~skiena/lydia/37720161.pdf

Lund, K., & Burgess, C. 1996. **Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence**. In the Behavior Research Methods, Instruments, and Computers, 28 (2). Pages 203-208.

http://www.springerlink.com/content/w06u365573x83884/

Martin-Bautista, M.J.; Larsen, H.L.; Nicolaisen, J.; Svendsen, T. 1997. **An Approach to an Adaptive Information Retrieval Agent using Genetic Algorithms with Fuzzy Set Genes.** In Proceeding of the 6th International Conference on Fuzzy Systems, Pages 1227-1232, Barcelona, Spain.

http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=619463

Marton, Y., Mohammad, S. and Resnik, P. 2009. **Estimating Semantic Distance Using Soft Semantic Constraints in Knowledge-Source–Corpus Hybrid Models.** In the Proceedings of the Empirical Methods in Natural Language Processing (EMNLP-2009), Pages 599-608, Singapore.

http://www.aclweb.org/anthology/D/D09/D09-1063

McDonald S. 2000. **Environmental Determinants of Lexical Processing Effort**. Ph.D. thesis, University of Edinburgh.

http://www.era.lib.ed.ac.uk/handle/1842/329

Mei, Qiaozhu and Ling, Xu and Wondra, Matthew and Su, Hang and Zhai, ChengXiang. 2007. **Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs**. In the Proceedings of the 16th International Conference on World Wide Web (WWW), Pages 171-180, Banff, Alberta, Canada.

http://www2007.org/papers/paper680.pdf

Melanie, Alt. 2008. **Emotional Responses To Color Associated With An Advertisement**. Master's Thesis, Graduate College of Bowling Green State University, Ohio.

http://etd.ohiolink.edu/send-pdf.cgi/Alt%20Melanie.pdf?bgsu1206377243

Mihalcea, R., Banea, C. and Wiebe, J. 2007. **Learning Multilingual Subjective Language via Cross-Lingual Projections**. In the Proceeding of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007). Pages 976-983, Prague, Czech Republic.

http://www.cs.pitt.edu/~wiebe/pubs/papers/acl07.pdf

Minsky, M. **The Emotion Machine**. Simon and Schuster, New York (2006).

http://www.amazon.com/Emotion-Machine-Commonsense-Artificial Intelligence/dp/0743276647/ref=ed_oe_p/

Mishne, Gilad and Rijke, Maarten de. 2006. **Moodviews: Tools for blog mood analysis**. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), Pages 153–154, 2006.

http://staff.science.uva.nl/~mdr/Publications/Files/aaai2006-mooddemo.pdf

Mohammad, Saif, Bonnie, D. and Graeme, H. 2008. **Computing Word-Pair Antonymy**. In the Proceeding of the Empirical Methods on Natural Language Processing (EMNLP-2008), Pages 982-991, Waikiki, Honolulu, Hawaii.

http://portal.acm.org/citation.cfm?id=1613715.1613843

Mohammad, Saif, and Turney, Peter. **Emotions Evoked by CommonWords and Phrases:Using Mechanical Turk to Create an Emotion Lexicon**. In the Proceeding of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, NAACL HLT 2010, Pages 26–34, Los Angeles, California.

http://aclweb.org/anthology/W/W10/W10-0204.pdf

Mohammad, Saif, Dorr S., and Dunne C. 2009. **Generating High-Coverage Semantic Orientation Lexicons from Overtly Marked Words and a Thesaurus.** In the Proceedings of EMNLP-2009, Pages 599-608.

www.aclweb.org/anthology/D09-1063

Moilanen, Karo; Pulman, Stephen and Zhang, Yue. 2010. **Packed Feelings and Ordered Sentiments: Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression**. In the Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010), ECAI 2010, Pages 36-43, Lisbon, Portugal.

www.cl.cam.ac.uk/~yz360/wassa10.pdf

Morinaga, Satoshi; Yamanishi, Kenji; Tateishi, Kenji and Fukushima, Toshikazu. 2002. **Mining Product Reputations on the Web**. In the Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Pages 341–349, Alberta, Canada.

www.ibis.t.u-tokyo.ac.jp/yamanishi/prm4.pdf

Mullen, T., and Collier, N. 2004. **Sentiment Analysis Using Support Vector Machines with Diverse Information Sources**. In Proceedings of the Empirical Methods in Natural Language Processing, Pages 412-418, Barcelona, Spain.

http://edu.tsuda.ac.jp/~mullen/Papers/emnlp_corrected.pdf

Narayan, D., Chakrabarti, D., Pande, P. and Bhattacharyya, P. 2002. **An Experience in Building the Indo WordNet - a WordNet for Hindi**. In the Proceeding of the 1st International Conference on Global WordNet, Mysore, India.

http://www.cfilt.iitb.ac.in/wordnet/webhwn/papers/gwn-2002.ps

Nasukawa, Tetsuya and Yi, Jeonghee. 2003. **Sentiment analysis: Capturing favorability using natural language processing**. In the Proceedings of the Conference on Knowledge Capture (K-CAP), Pages 70-77, Sanibel Island, Florida, USA.

http://dl.acm.org/citation.cfm?id=945658

Nigam, K., and Hurst, M. 2004. **Towards A Robust Metric of Opinion**. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text.

www.kamalnigam.com/papers/metric-EAAT04.pdf

Nivre, J. 2009. ***Non-Projective Dependency Parsing in Expected Linear Time***. In Proceeding of the ACL IJCNLP.

www.aclweb.org/anthology/P/P09/P09-1040.pdf

Bibliography

Nivre J. and Nilsson J. 2005. **Pseudo Projective Dependency Parsing**. In the Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), Pages 99–106.

http://w3.msi.vxu.se/~nivre/papers/acl05.pdf

Nivre J., Hall J. and Nilsson J. 2006. **MaltParser: A Data- Driven Parser-Generator for Dependency Parsing**. In the Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006), Pages 2216-2219, Genoa, Italy.

www.lrec-conf.org/proceedings/lrec2006/pdf/162_pdf.pdf

Nivre, J., Hall, J., Kubler, McDonald, S., R., Nilsson, J., Riedel, S. and Yuret, D. 2007. **The CoNLL 2007 Shared Task on Dependency Parsing**. In Proceeding of the EMNLP/CoNLL-2007.

http://acl.ldc.upenn.edu/D/D07/D07-1096.pdf

Ohana, Bruno and Tierney, Brendan. 2009. **Sentiment classification of reviews using SentiWordNet**. In the Proceeding of the 9th IT&T Conference, Dublin, Ireland. http://arrow.dit.ie/ittpapnin/13/

Palmer Martha, Bhatt Rajesh, Narasimhan Bhuvana, Rambow Owen, Sharma Dipti Misra and Xia Fei. **Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure.** In the Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Pages 259-268.

http://ltrc.iiit.ac.in/icon_archives/ICON2009/Papers/pdf/28.pdf

Palmer Martha, Gildea Dan and Kingsbury Paul. 2005. **The Proposition Bank: A Corpus Annotated with Semantic Roles**. In Computational Linguistics Journal, 31:1.

www.cs.rochester.edu/~gildea/palmer-propbank-cl.pdf

Pang, B., Lee, L. and Vaithyanathan, S. 2002. **Thumbs up? Sentiment Classification using Machine Learning Techniques**. In the Proceedings of the Empirical Methods on Natural Language Processing (EMNLP 2002), Pages 79-86, Pennsylvania, USA.

http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

Pang, Bo and Lee, Lillian Lee. 2005. **Seeing Stars: Exploiting Class Relationships for Sentiment Categorization With Respect To Rating Scales**. In the Proceedings of the 43rd Annual Meeting of the ACL, Pages 115-124, Ann Arbor, USA.

www.cs.cornell.edu/home/llee/papers/pang-lee-stars.pdf

Pang, Bo and Lee, Lillian. 2004. **A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts**. In the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004), Barcelona, Spain.

http://acl.ldc.upenn.edu/P/P04/P04-1035.pdf

Parton Kristen, McKeown Kathleen, Coyne Bob, Diab Mona, Grishman Ralph, Hakka-ni-Tür Dilek, Harper Mary, Ji Heng, Ma Wei Yun, Meyers Adam, Stolbach Sara, Sun Ang, Tur Gok-han, Xu Wei and Yaman Sibel. 2009. **Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task**. In the Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Pages 423–431, Suntec, Singapore.

www.aclweb.org/anthology/P09-1048

Picard, R. 1997. **Affective Computing**. MIT Press, Cambridge.

http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=4062

Plutchik, R. **The Nature of Emotions**. 2001. American Scientist 89(4), Pages 344–350.

Polanyi, Livia; Culy, Chris; Van den Berg, Martin; Thione, Gian Lorenzo and Ahn, David. 2004. **Sentential Structure and Discourse Parsing**. In the Proceeding of the Workshop on Discourse Annotation ACL2004, Pages 80-87, Barcelona, Spain.

www.fxpal.com/publications/FXPAL-PR-04-285.pdf

Quirk, Randolph, Greenbaum, Sidney, Leech Geoffrey and Svartvik, Jan. 1985. **A comprehensive grammar of the English language**.

http://www.amazon.com/Comprehensive-Grammar-English-Language/dp/0582517346

Read, Jonathon. 2005. **Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification.** In the Proceedings of the Student Research Workshop, ACL 2005, Pages 43–48, Ann Arbor, USA.

http://portal.acm.org/citation.cfm?id=1628969

Riloff, E., Wiebe, J., and Wilson, T. 2003. **Learning Subjective Nouns Using Extraction Pattern Bootstrapping**. In Proceedings of the Seventh Conference on Natural Language Learning Conference, Pages25-32, Edmonton, Canada.

www.cs.utah.edu/~riloff/pdfs/conll03.pdf

Robkop, K., Thoongsup, S., Charoenporn, T., Sornlertlamvanich, V. and Isahara, H. 2010. **WNMS: Connecting the Distributed WordNet in the Case of Asian WordNet**. In the Proceeding of 5th International Conference of the Global WordNet Association (GWC-2010), Mumbai, India.

http://www.cfilt.iitb.ac.in/gwc2010/pdfs/68_WNMS_WordNet__Robkop.pdf

Ruhl, C. 1989. **On Monosemy**. New York: SUNY Press.

http://www.sunypress.edu/p-594-on-monosemy.aspx

Saha, G.K.; Debnath, A.B. S. 2004. **Computer Assisted Bangla POS Tagging**. iSTRAN, Tata McGraw-Hill, NewDelhi.

http://www.getcited.org/pub/103446371

Salton, G., Wong, A., & Yang, C.-S. 1975. **A Vector Space Model for Automatic Indexing**. In Communications of the ACM, 18 (11), Pages 613-620.

http://openlib.org/home/krichel/courses/lis618/readings/salton75.pdf

Salvetti, Franco; Lewis, Stephen and Reichenbach, Christoph. 2004. **Automatic Opinion Polarity Classification of Movie Reviews**. In Colorado Research in Linguistics, Volume 17, Issue 1.

http://www.colorado.edu/ling/CRIL/Volume17_Issue1/paper_SALVETTI_LEWIS_REICHENBACH.pdf

Sarkar, Sandipan and Bandyopadhyay, Sivaji. 2009. **Study on Rule-Based Stemming Patterns and Issues in a Bengali Short Story-Based Corpus**. In the Proceeding of the International Conference of the Natural Language Processing (ICON 2009), Pages 346-351, Hyderabad, India.

Seeker Wolfgang, Bermingham Adam, Foster Jennifer and Hogan Deirdre. 2009. **Exploiting Syntax in Sentiment Polarity Classification**. National Centre for Language Technology Dublin City University, Ireland.

www.nclt.dcu.ie/slides/wolfgangS.pdf

Seki, Yohei, Eguchi, Koji, and Kando, Noriko. 2004. **Analysis of Multi-Document Viewpoint Summarization Using Multi-Dimensional Genres**. In the Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Pages 142–145, Menlo Park, CA.

http://aaaipress.org/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-026.pdf

Shneiderman, Ben. 1992 **Tree Visualization with Treemaps: A 2nd Space-Filling Approach**. In ACM Transactions on Graphics, 11(1), Pages 92-99.

http://dl.acm.org/citation.cfm?id=115768

Singh, Push. 2002. ***The public acquisition of commonsense knowledge.*** In the Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access, Palo Alto, CA.

http://web.media.mit.edu/~push/AAAI2002-Spring.pdf

Snyder, Benjamin and Barzilay, Regina. 2007. **Multiple Aspect Ranking using the Good Grief Algorithm**. In the Proceedings of the NAACL HLT 2007, Pages 300–307.

http://pages.cs.wisc.edu/~bsnyder/presentations/naacl07.pdf

Somasundaran, Swapna. 2010. **Discourse-level relations for Opinion Analysis**. PhD Thesis, University of Pittsburgh.

www.cs.pitt.edu/~swapna/papers/somasundaranThesis.pdf

Somasundaran, Swapna; Namata, Galileo; Wiebe, Janyce and Getoor, Lise. 2009. **Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification.** In the EMNLP 2009, Pages 170-179, Singapore.

www.somasundaran.net/papers/SomasundaranEtal-emnlp2009.pdf

Speer, R., Havasi, C. and Lieberman, H. 2008. **Analogy Space: Reducing the Dimensionality of Common Sense Knowledge.** In the Proceeding of the Proceedings of the 23rd national conference on Artificial intelligence (AAAI 2008), Pages 548-553, Chicago, Illinois.

http://analogyspace.media.mit.edu/media/speerhavasi.pdf

Speriosu, Michael and Sudan, Nikita and Upadhyay, Sid and Baldridge, Jason. 2011. **Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph**. In the Proceedings of the 1st Workshop on Unsupervised Learning in NLP, Pages 53–63, Edinburgh, Scotland.

http://aclweb.org/anthology-new/W/W11/W11-2207.pdf

Srinivas, M. and Patnaik, L. M. 1994. **Adaptive Probabilities of Crossover and Mutation in Genetic algorithms**. In the IEEE Transactions on Systems, Man and Cybernetics, 24(4): Pages 656–667.

http://eprints.iisc.ernet.in/6971/2/adaptive.pdf

Stone, Philip J. 1966. **The General Inquirer: A Computer Approach to Content Analysis.** The MIT Press, 1966.

http://www.wjh.harvard.edu/~inquirer/

Stone, P.J. 1997 **Thematic Text Analysis: New Agendas for Analyzing Text Content**. Chapter 2, In Text Analysis for the Social Science, Lawrence Erlbaum Associates, Publishers.

http://books.google.co.in/books?id=1-tg1tHuUjAC&pg=PA35&lpg=PA35&dq=Thematic+text+analysis:+new+agendas+for+analyzing+text+content&source=bl&ots=GpV2dSPUll&sig=_vyjsiR7o3nrmhlFee01FRYSh4A&hl=en&ei=eYrLTv_kD5GsrAfIw8WxDA&sa=X&oi=book_result&ct=result&resnum=3&ved=0CDAQ6AEwAg#v=onepage&q=Thematic%20text%20analysis%3A%20new%20agendas%20for%20analyzing%20text%20content&f=false

Strapparava, C. and Mihalcea, R. 2008. **Learning To Identify Emotions in Text.** In the Proceedings of the ACM symposium on Applied computing, Pages 1556-1560, New York, USA.

www.cse.unt.edu/~rada/papers/strapparava.acm08.pdf

Strapparava, Carlo and Ozbal, Gozde. 2010. **The Color of Emotions in Texts**. In Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (COGALEX II), COLING 2010, Pages 28-32. Beijing, China.

http://www.aclweb.org/anthology/W10-3405

Strapparava, Carlo and Valitutti, Alessandro. 2004. **WordNet-Affect: an affective extension of WordNet.** In the Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Pages 1083-1086, Lisbon.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.4281&rep=rep1&type=pdf

Subasic, P., and Huettner, A. 2001. **Affect Analysis of Text Using Fuzzy Semantic Typing**. In the IEEE Transactions on Fuzzy Systems 9, 4, Pages 483-496.

www.ieeexplore.ieee.org/iel5/91/20371/00940962.pdf

Taboada, M., Anthony, C. and Voll, K. 2006. **Methods for Creating Semantic Orientation Dictionaries**. In the Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Pages 427-432, Genoa, Italy.

http://www.sfu.ca/~mtaboada/docs/Taboada_et_al_LREC_2006.pdf

Takamura Hiroya, Inui Takashi and Okumura Manabu. 2005. **Extracting Semantic Orientations of Words using Spin Model**. In the Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005), Pages 133-140.

http://acl.ldc.upenn.edu/P/P05/P05-1017.pdf

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. 2010. **Sentiment Strength Detection In Short Informal Text**. In the Journal of the American Society for Information Science and Technology, 61(12), Pages 2544–2558.

www.scit.wlv.ac.uk/~cm1993/papers/SentiStrengthPreprint.doc

Tong, R.M. 2001. **An Operational System for Detecting and Tracking Opinions in On-Line Discussions**. Working Notes in the Proceeding of the Workshop on Operational Text Classification, ACM SIGIR 2001, Pages 1-6, New York, USA.

http://www.daviddlewis.com/events/otc2001/presentations/otc01-tong-paper.pdf

Torii, Yoshimitsu; Das, Dipankar; Bandyopadhyay, Sivaji and Okumura, Manabu. 2011. **Developing Japanese WordNet Affect for Analyzing Emotions**. In the proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011). ACL-HLT 2011, Pages 80-86, Portland, Oregon, USA.

www.aclweb.org/anthology/W11-1710

Turney Peter D. and Littman Michael L. 2003. **Measuring Praise and Criticism: Inference of Semantic Orientation from Association**. In ACM Transactions on Information Systems (TOIS), 21(4), Pages 315–346.

http://arxiv.org/ftp/cs/papers/0309/0309034.pdf

Turney, P. 2002. **Thumbs up or thumbs down? Semantic orientation Applied to Unsupervised Classification of Reviews**. In the Proceeding of the Association for Computational Linguistics (ACL-2002), Pages 417-424, Philadelphia, Pennsylvania.

http://www.ldc.upenn.edu/acl/P/P02/P02-1053.pdf

Umansky-Pesin, Shulamit and Reichart, Roi and Rappoport, Ari. 2010. **A Multi-Domain Web-Based Algorithm for POS Tagging of Unknown Words**. In COLING 2010, Pages 1274-1282, Beijing, China.

http://www.cs.huji.ac.il/~arir/10-pos-unknown-words-coling-2010.pdf

Vaidya Ashwini, Husain Samar, Mannem Prashanth, and Sharma Dipti Misra. 2009. **A Karaka Based Annotation Scheme for English**. In the Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science; Vol. 5449. Pages: 41 – 52.

http://researchweb.iiit.ac.in/~prashanth/papers/vaidya-husain-mannem-sharma-cicling09.pdf

Valitutti, Alessandro, Strapparava, Carlo and Stock, Oliviero. 2004. **Developing Affective Lexical Resources**. In PsychNology Journal, 2004 Volume 2, Pages 61 – 83.

http://www.psychnology.org/File/PSYCHNOLOGY_JOURNAL_2_1_VALITUTTI.pdf

Von Ahn, Luis and Dabbish, Laura. 2004. **Labeling Images with a Computer Game**. In the Proceeding of SIGCHI conference on Human factors in Computing Systems, CHI 2004, Pages 319-326, Vienna, Austria.

http://doi.acm.org/10.1145/985692.985733

Wainer, H. 1997. **A Rose by Another Name**. *Visual Revelations, Copernicus Books*, New York.

http://books.google.co.in/books?id=NcEb3dxbnrsC&pg=PR7&lpg=PR7&dq=A+Rose+by+Another+Name%2BWainer&source=bl&ots=E_ntWXbpiV&sig=K1y0V5BdEN_AIqpKDhnYf5ieFyg&hl=en&ei=B1HeTpTnNMyqrAe1laDUCA&sa=X&oi=book_result&ct=result&resnum=1&ved=0CCUQ6AEwAA#v=onepage&q&f=false

Wan, Xiaojun. 2008. **Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis**. *In EMNLP-2008*, Pages 553-561.

http://www.aclweb.org/anthology-new/D/D08/D08-1058.pdf

Wiebe, J., Bruce, Rebecca F. and O'Hara Thomas P. 1999. **Development and Use of a Gold Standard Data Set for Subjectivity Classifications**. In the Proceedings of the Association for Computational Linguistics (ACL 1999), Pages 246–253, College Park, Maryland.

http://acl.ldc.upenn.edu/P/P99/P99-1032.pdf

Wiebe, Janyce and Mihalcea, Rada. 2006. **Word Sense and Subjectivity**. In the Proceeding of COLING/ACL-06, Pages 1065-1072, Sydney, Australia.

http://www.cs.pitt.edu/~wiebe/pubs/papers/acl06.pdf

Wiebe, Janyce and Rapaport, William. 1988. **A Computational Theory of Perspective and Reference in Narrative**. In the Proceedings of the Association for Computational Linguistics (ACL 1988), Pages 131–138, Buffalo, New York.

http://acl.ldc.upenn.edu/P/P88/P88-1016.pdf

Wiebe, Janyce and Riloff, Ellen. 2005. **Creating Subjective and Objective Sentence Classifiers From Unannotated Texts**. In the Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005), Pages 475-486, Mexico.

http://www.cs.pitt.edu/~wiebe/pubs/papers/cicling05.pdf

Wiebe, Janyce. 1990. **Recognizing Subjective Sentences: A Computational Investigation of Narrative Text**. Ph.D. dissertation, Technical Report 90-03, Buffalo: SUNY Buffalo Department of Computer Science.

Available on Request from the Author.

http://dl.acm.org/citation.cfm?id=101562

Wiebe, Janyce. 2000. **Learning Subjective Adjectives from Corpora**. In the Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, CA: AAAI 2000, Pages 735-740, Menlo Park.

http://www.cs.columbia.edu/~vh/courses/LexicalSemantics/Orientation/wiebe-aaai2000.pdf

Wilks, Yorick and Bien, Janusz. 1983. **Beliefs, Points of View and Multiple Environments**. In the Proceeding of the International NATO symposium on Artificial and Human Intelligence, Pages 147-171, Lyon, France.

http://csjarchive.cogsci.rpi.edu/1983v07/i02/p0095p0119/MAIN.PDF

Wilson, T., Wiebe, J. and Hoffmann, P. 2005. **Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis**. In the Proceeding of Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP-2005), Pages 347-354, Stroudsburg, PA, USA.

http://www.cs.pitt.edu/~wiebe/pubs/papers/emnlp05polarity.pdf

Wilson, Theresa; Wiebe, Janyce and Hwa, R. 2006. **Recognizing Strong and Weak Opinion Clauses**. In Computational Intelligence, 22(2), Pages 73-99.

http://club.fom.ru/books/ci06.pdf

Whitelaw, C., Garg, N., and Argamon, S. 2005. **Using Appraisal Groups For Sentiment Analysis**. In Proceedings of the 14th ACM Conference on Information and Knowledge Management, Pages 625-631.

http://lingcog.iit.edu/doc/appraisal_sentiment_cikm.pdf

Yi, Jeonghee and Nasukawa, Tetsuya and Bunescu, Razvan and Niblack, Wayne. 2003. **Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques**. In the Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM), Pages 427-434, Washington, DC, USA.

http://www.cs.utexas.edu/~razvan/papers/icdm2003.pdf

Yi, Jeonghee and Niblack, Wayne.2005. **Sentiment mining in WebFountain**. In the Proceedings of the International Conference on Data Engineering (ICDE), Pages 1073-1083.

http://suraj.lums.edu.pk/~cs631s05/Papers/sentiment_webfountain.pdf

Yu, H. and Hatzivassiloglou, V. 2003. **Towards Answering Opinion Questions: Separating Facts From Opinions And Identifying The Polarity Of Opinion Sentences**. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Pages 129-136.

www.ldc.upenn.edu/acl/W/W03/W03-1017.pdf

Zhao, Jun; Liu, Kang; Wang, Gen. 2008. **Adding Redundant Features for CRFs-based Sentence Sentiment Classification**. *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2008), Pages 117–126, Honolulu.

http://www.aclweb.org/anthology-new/D/D08/D08-1013.pdf

Zhou Liang and Hovy Eduard. 2006. **On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs**. In the Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, Stanford, CA.

http://www.isi.edu/natural-language/people/hovy/papers/06SprSymp-weblog.pdf

Zhuang, Li and Jing, Feng and Zhu, Xiao-Yan. 2006. **Movie Review Mining and Summarization**. In the Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM), Pages 43-50, Arlington, Virginia, USA.

http://research.microsoft.com/en-us/um/people/leizhang/Paper/cikm06_movie.pdf