Dr Sentiment Creates SentiWordNet(s) for Indian Languages Involving Internet Population

Amitava Das and **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering
Jadavpur University
amitava.santu@gmail.com, sivaji_cse_ju@yahoo.com

Abstract

A typical computational approach to sentiment analysis starts with prior polarity lexicons where entries are tagged with their prior out of context polarity as perceived by human beings based on their cognitive knowledge. Till date, most of the research efforts found in sentiment analysis literature deal with English texts. In this paper, we propose an interactive gaming (Dr Sentiment) technology to create and validate SentiWordNet(s) for three Indian languages, Bengali, Hindi and Telugu by involving Internet population. A number of automatic, semiautomatic and manual validations and evaluation methodologies have been adopted to measure the coverage and credibility of the developed SentiWordNet(s).

1 Prior Polarity Lexicon

In order to identify sentiment from a text, lexical analysis plays a crucial role. For example, words like ***love, hate, good*** and ***favorite*** directly indicate sentiment or opinion. Various previous works (Pang et al., 2002; Wiebe and Mihalcea, 2006; Esuli et. al., 2006) have already proposed techniques for making dictionaries for those sentiment words. But identification of polarity orientation of those words is another vital research issue.

Polarity Identification and classification of such sentiment lexicons is a hard contextual semantic disambiguation problem. The regulating aspects of semantic orientation of a lexicon are natural language context information (Pang et al., 2002), language properties (Wiebe and Mihalcea, 2006), domain pragmatic knowledge (Aue and Gamon, 2005) and the most challenging time dimension (Read, 2005).

The following two examples show that the polarity tag associated with a sentiment word depends on context / domain knowledge and time dimension respectively.

Example 1: I prefer Limuzin as it is **longer** than Mercedes.

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

In the previous two example sentences the word **long** has been used as a sentiment/opinion word. In the first sentence the word long depicts positive sentiment whereas in the second example it expresses a negative sentiment.

Example 2: 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 previous example shows that the polarity tag associated with a sentiment word depends on the time dimension.

Therefore lexicon level polarity assignment is a bit difficult. Previous researchers (Wiebe and Mihalcea, 2006; Aue and Gamon, 2005) have proposed corpus heuristic based polarity assignment at lexicon level. The ratios of the word distribution as positive or negative and the total occurrence of a particular word in a domain corpus are calculated. Suppose total occurrence of a word "**long**" in a domain corpus is n. The positive and negative occurrence of that word is S_p and S_n respectively. Therefore in a developed sentiment lexicon the assigned positivity and negativity score of that word will be as follows:

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

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

These associative scores are called **prior polarity**. Prior polarity is an approximation value and not exact. Prior polarity sentiment lexicons are necessary for any new language as a foundation to start the exploration of computational sentiment analysis for the language. Although contextual polarity disambiguation techniques are still required for further sentiment/opinion analysis task. Sentiment lexicons only provide a good baseline, i.e., dictionary based approach without using any NLP techniques. The performance of the polarity classifier has been reported in the Section 6.1. Feature ablation method, reported in Table 6 shows that only dictionary based approach give good baseline score.

2 Motivations

Several prior polarity sentiment lexicons are available for English such as SentiWordNet (Esuli et. al., 2006), Subjectivity Word List (Wilson et. al., 2005), WordNet Affect list (Strapparava et al., 2004), Taboada's adjective list (Voll et al., 2006).

Among these publicly available sentiment lexicon resources we find that SentiWordNet is most widely used (number of citation is higher than other resources[1]) in several applications such as sentiment analysis, opinion mining and emotion analysis. SentiWordNet is an automatically constructed lexical resource for English that assigns a positivity score and a negativity score to each WordNet synset. Therefore we decided to develop SentiWordNet for new languages.

A number of research endeavors could be found in literature for creation of Sentiment Lexicon in several languages and domains. These techniques can be broadly categorized in two genres, one follows classical manual annotation (Andreevskaia and Bergler, 2006);(Wiebe and Riloff, 2006); (Mohammad et al., 2008) techniques and the other proposes various automatic techniques (Tong, 2001). Both types of techniques have few limitations.

Manual annotation techniques are undoubtedly trustable but it generally takes time. A large number of annotators are needed to balance one's sentimentality in order to reach agreement. On the other hand, automatic techniques demands manual validations and are dependent on the corpus availability in the respective domain.

Polarity of sentiment lexicons depends on multiple factors such as: language specific, domain specific, time specific and may be other hidden multiple aspects. Moreover sentiment is a social understanding which we, the human being learn from the society through everyday cognitive interactions. Therefore sentiment is one's or more than one's out of context (as prior polarity has no contextuality) psychology regarding any topic or concept.

Therefore involving people is the best way to capture the sentiment of the human society. But engaging human annotators is not an easy task. Hence we have developed an online game to attract internet population for the creation of SentiWordNet(s) automatically. Involvement of Internet population is an effective approach as the population is very high in number and ever growing (approx. 360,985,492)[2], there are peoples with various languages, cultures, age etc. Internet population is thus not biased towards any domain, language or particular society.

The developed online game "***Dr Sentiment***", revolutionize the idea of making prior polarity sentiment lexicon for any new language by involving internet population. We have evaluated the coverage and credibility of the generated sentiment lexicons by Dr Sentiment.

3 Source Lexicon Acquisition

SentiWordNet and Subjectivity Word List have been identified as the most reliable source lexicons. A merged sentiment lexicon has been developed from both the resources by removing the duplicates. It has been observed that 64% of the single word entries are common in the Subjectivity Word List and SentiWordNet. The new merged sentiment lexicon consists of 14,135 numbers of tokens. Several filtering techniques have been applied to generate the new list.

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 value 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 (Rada et al., 2007) from the Subjectivity word list.

In the next stage the words whose POS category in the Subjectivity word list is undefined and tagged as "***anypos***" are discarded. These words may generate sense ambiguity issues in the next stages of subjectivity detection.

Some words in the Subjectivity word list are inflected e.g., memories. These words would be stemmed during the translation process, but some

[1] http://citeseerx.ist.psu.edu/

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

Table 1: English SentiWordNet and Subjectivity Word List Statistics

Words	SentiWordNet		Subjectivity Word List	
Entries	Single	Multi-Word	Single	Multi-Word
	115424	79091	5866	990
Unambiguous	20789	30000	4745	963
Ambiguous	Threshold	Orientation Strength	Subjectivity Strength	POS
	86944	30000	2652	928

words present no subjectivity property after stemming (memory has no subjectivity property). A word may occur in the subjectivity list in many inflected forms. Individual clusters for the words that share the same root form are created and then 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. Various statistics of the English SentiWordNet and Subjectivity Word List are reported in Table 1.

4 Dr Sentiment

There are several motivations behind developing an intuitive game to automatically create multilingual SentiWordNet(s). Sentiment lexicon generation from any source language to target language has several issues or limitations:

- Source language word may have no sentiment value in target language (cross language limitation)
- Sentiment scores of the source and the target languages may not be equal
- Relative sentiment score is needed rather than absolute score
- Language / Culture specific lexicons should be included
- Sentiment score should be updated with time

In the history of Information Retrieval research there is a milestone when ESP[3] game (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. We are highly motivated by the success of the Image Labeler game and thus proposed an intuitive game to create and validate sentiment lexicons for three Indian languages involving the large number of Internet population. A brief statistics about Internet population has been reported in Table 2.

Dr Sentiment is an interactive game[4]. It asks each player a set of simple questions and reveals his/her sentimental status. The lexicons tagged by this system are credible as it is tagged by human beings. Moreover, all the aspects of limitation of lexicon development have been covered in this strategy. As each player can play in their native language so there is no issue of cross language limitations. Different tables are maintained for different languages. Each player assigns the sentiment tag to a word as suggested by Dr. Sentiment as part of question type 1 (Q1). Relative sentiment score has been calculated by question type 2 (Q2) described in Section 4.1.2. Language or culture specific words are being captured by question types 3 (Q3) and 4 (Q4) as described in Sections 4.1.3 and 4.1.4 respectively). It is not a static sentiment lexicon set as it is updated regularly. Almost 100 players per day are currently playing it throughout the world in different languages. Snapshots of different screens from the game are presented in Figure 1.

For word based translation Google translation[5] service has been used. It is a nice web service that translates at least at word level with very less ambiguity. To make the gaming interface more interesting images has been added. These images have been retrieved by Google image search API and to avoid biasness we have randomized among the first ten images retrieved by Google.

4.1 Strategy

There are four types of questions: Q1, Q2, Q3 and Q4. 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. The questions are randomly asked to keep the game more interesting.

[3] http://www.espgame.org/

[4] http://www.amitavadas.com/Sentiment%20Game/

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

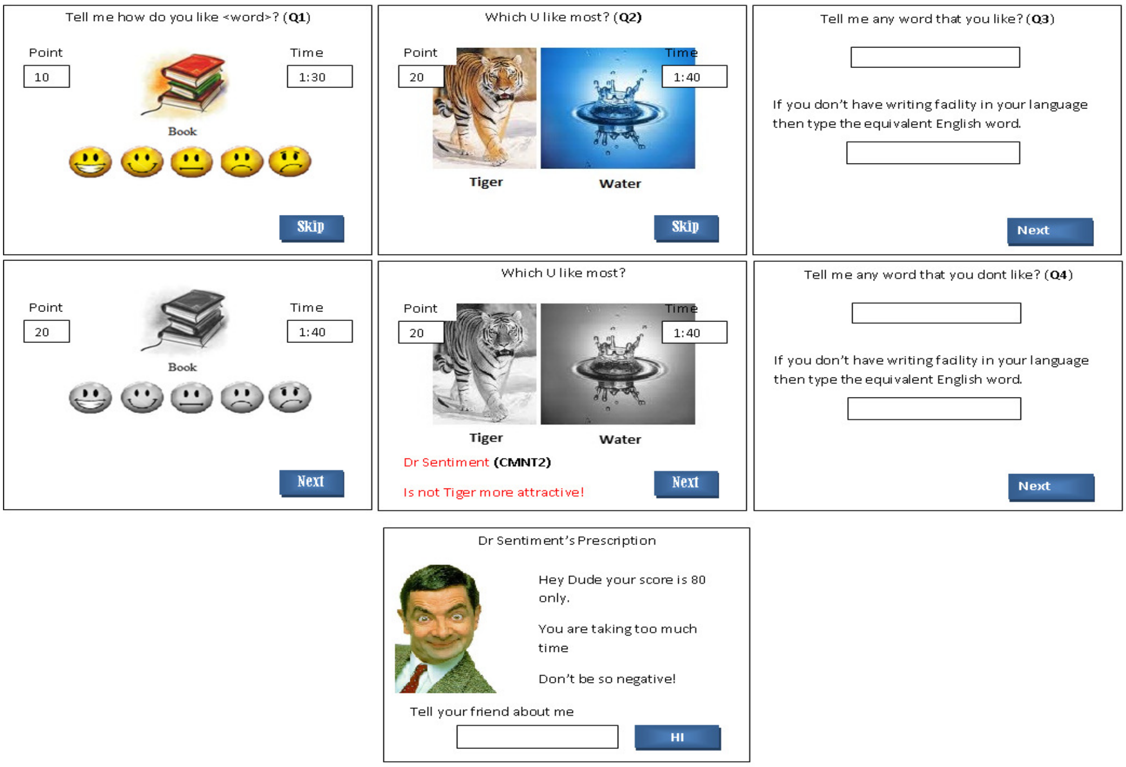


Figure 1: Snapshots from Dr Sentiment Game

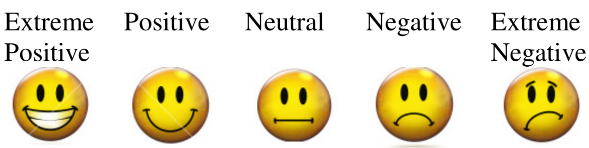
Table 2: Internet Usage and Population Statistics

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 %

4.1.1 Q1

An English word from the English SentiWordNet and the Subjectivity Word List is randomly chosen. The 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. The words are shown in the player's own language as specified by the player in the login page.

The sentiment scores are calculated based on the different emoticons pressed by different players and the sentiment score is scaled 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).



Each word as tagged by each player along with its associate property (POS, Offset) are inserted into the language specific tables. The Offset property refers to the English WordNet Synset offset of the word. The calculated positivity and negativity scores are stored accordingly.

4.1.2 Q2

Randomly n (presently 2-4) words have been chosen from the source English lexicon along with their images as retrieved by Google API.

Each player is asked to select one of them that he/she likes most. If the selection by the player matches the English SentiWordNet then the positive sentiment score of the selected word is mapped into the target language. Otherwise, the average of the differences in positive sentiment scores of any two words in the question word set is added to the positive sentiment score of the word selected by the player. The relative score is stored in the corresponding language table.

4.1.3 Q3

The player is asked for any positive word in the language of the player. The words are added to the corresponding language table with pos: 0.5 and neg: 0.0 score.

4.1.4 Q4

A player is asked by Dr Sentiment about any negative word in the language of the player. The word is then added to the corresponding language table with pos: 0.0 and neg: 0.5 score.

4.2 Comment Architecture

There are three types of Comments, Comment type 1 (CMNT1), Comment type 2 (CMNT2) and the final comment as Dr Sentiment's prescription. CMNT1 type and CMNT2 type comments are associated with question types Q1 and Q2 respectively.

4.2.1 CMNT1

Comment type 1 has 5 variations as shown in the Comment table in Table 3.

- 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 comment type table according to their category.

4.2.2 CMNT2

The strategy here is as same as the CMNT 1. Comment type 2 has only 2 variations as.

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

4.3 Dr Sentiment's Prescription

The final prescription depends on various factors such as total number of positive, negative or neutral comments and the total time taken by any player. The final prescription also depends on the range of the values of the above factors. More than one prescription may be identified based on the above factors and the final prescription is selected randomly.

5 Senti-Mentality

Several analyses have been done on the developed sentiment lexicons to understand the sentimental behavior of people depending upon location, age, sex, profession and etc. The login form of the "**Dr Sentiment**" game asks the player to provide several information such as country, city, age, sex, profession etc. A tracking system keeps track of every player's tagged words. A word previously tagged by a player is avoided by the tracking system for the next time playing as our intension is to tag more and more words involving Internet population.

Table 3: Comments

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 hiding from me? I am Dr Sentiment.

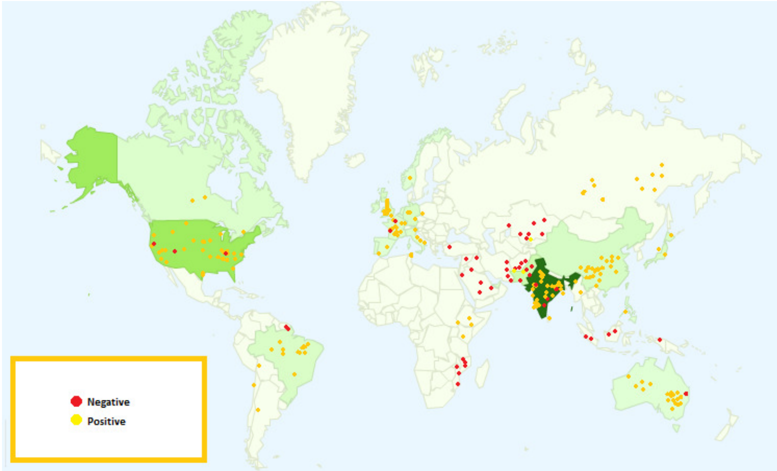


Figure 2: Geospatial Senti-Mentality

We observe 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. Statistical analyses reveal some interesting data as described below.

5.1 Concept-Culture-Wise Analysis

During analysis we found an interesting outcome. The word "**blue**" get 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. The graphical illustration in Figure 2 explains the situation. The observation is that most of the negative tags are coming from the middle-east and especially from the Islamic countries. We found a line in Wiki[6] (see in Religion Section) that may give 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 other explanations may be there for this interesting observation for sentiment lexicon creation.

5.2 Age-Wise Analysis

Another interesting observation is that sentimentality may vary age-wise. For better understanding we look at the total statistics and the age wise distribution of all the players. Total 533 players have taken part till date. The total number of players for each range of age are shown at top of every bar. In the Figure 3 the horizontal bars are divided into two colors (Green depicts the Positivity and Red depicts the negativity) according to the total positivity and negativity scores, gathered during playing. This sociological study gives an idea that variation of sentimentality with age.

5.3 Gender Specific

It is observed from the statistics collected that women are more positive than a man. The variations in sentimentality among men and women are shown in the following Figure 4.

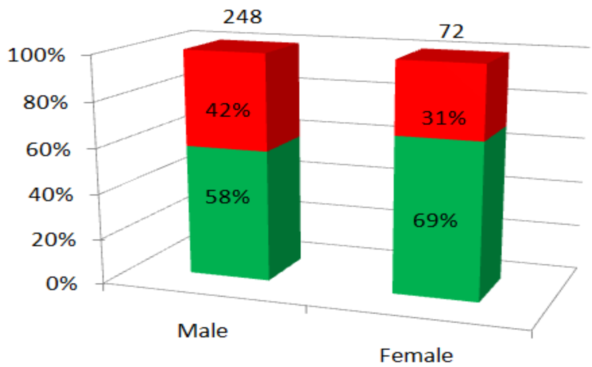


Figure 4: Gender Specific Senti-Mentality

5.4 Other-Wise

We have described several important observations in the previous sections and there are other important observations as well. Some of the important dimensions may be country, city, profession etc. Studies on the combinations of these dimensions, such as, location-age, location-profession and language-location may reveal some interesting results.

6 Expected Impact of the Resources

The Google translation services may produce wrong translation but the same erroneous output then what will be the impact into the targeted language-specific SentiWordNet(s)?

We have manually checked Google word-level translation for Indian languages and there were

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

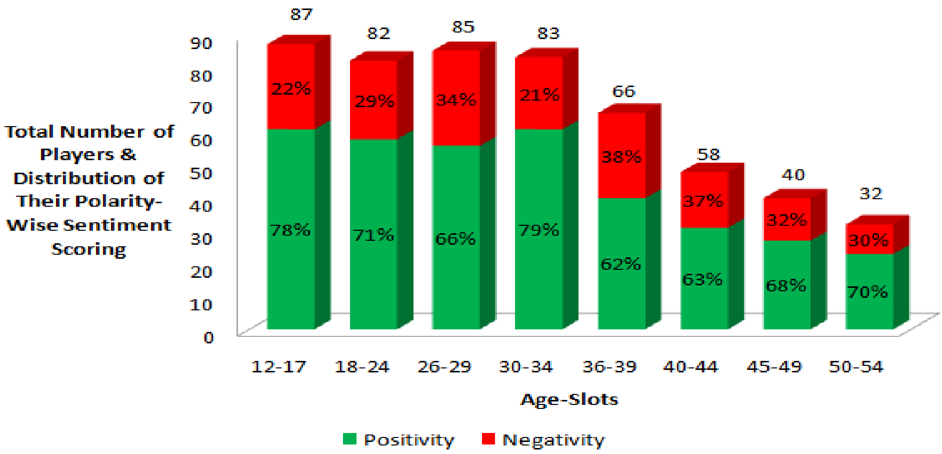


Figure 3: Age-Wise Senti-Mentality

will be generated every time. So the same erroneous output of any source word gets tagged by players of the target language. The background database of the system stores data into language specific tables, so there is no inter-language ambiguity. The erroneous outputs by Google may introduce errors for cross-lingual use but still the developed SentiWordNet(s) are useful for monolingual use. Presently Bengali SentiWordNet contains 20,546, Hindi SentiWordNet contains 13,889 and Telugu SentiWordNet contains 10,204 unique entries.

The generated lexicons are important resources for three languages for sentiment/opinion or emotion analysis task. Moreover the other non linguistic dimensions are very much important for further analysis and 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.

7 Evaluation

The inventors of the SentiWordNet Andera Esuli and Fabrizio Sebastiani (Esuli and Fabrizio, 2006) have calculated the reliability of the sentiment scores attached to each synsets in the SentiWordNet. They have tagged sentiment words in the English WordNet with positive and negative sentiment scores. We have proposed two extrinsic evaluation strategies for the developed Bengali SentiWordNet based on the two usages of the sentiment lexicon, subjectivity classifier and polarity identifier. The Hindi and Telugu SentiWordNet(s) have been partly evaluated. SentiWordNet(s) for other languages have not been evaluated yet. We plan to do that in future.

7.1 Coverage

We experimented with NEWS and BLOG corpora for subjectivity detection. Sentiment lexicons are generally domain independent but 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). The statistics of the NEWS and BLOG corpora is reported in Table 4.

Table 4: Bengali Corpus Statistics

	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 5: Subjectivity Classifier using SentiWordNet

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%**

For comparison with the coverage of English SentiWordNet the same subjectivity classifier (Das and Bandyopadhyay, 2009) has been applied on Multi Perspective Question Answering (MPQA) (NEWS) and IMDB Movie review corpus along with English SentiWordNet.

The result of the subjectivity classifier on both the corpus proves that the coverage of the Bengali SentiWordNet is reasonably good.

Table 6: Polarity Performance Using Bengali SentiWordNet

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%

The subjectivity word list used in the subjectivity classifier is developed from the IMDB corpus and hence the experiments on the IMDB corpus have yielded high precision and recall scores. The developed Bengali SentiWordNet is domain independent and still its coverage is very good as shown in Table 5. We are not able to evaluate the Hindi and Telugu SentiWordNet(s) as there is no publicly available annotated data for these languages.

7.2 Credibility of Polarity Scores

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 these lexicons, 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) has been developed using the Bengali SentiWordNet along with some other linguistic features. Feature ablation method proves that the generated SentiWordNet gives a good baseline. Although contextual polarity disambiguation techniques are required that may use multiple features.

Feature ablation method proves that the associated polarity scores in the developed Bengali SentiWordNet are reliable. Table 6 shows the performance of a polarity classifier using the Bengali SentiWordNet. The polarity wise overall performance of the polarity classifier is reported in Table 7.

Table 7: Polarity-wise Performance Using Bengali SentiWordNet

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

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

Table 8: Evaluation of Polarity Score of Developed Hindi SentiWordNet

Polarity	Positive	Negative
Percentage	88.0%	91.0%

An arbitrary 300 words have been chosen from the Hindi SentiWordNet for human evaluation. Two persons are asked to manually check it and the result is reported in Table 8. The coverage of the Hindi SentiWordNet has not been evaluated, as no manually annotated sentiment corpus is available.

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 9. The coverage of Telugu SentiWordNet has not been evaluated, as no manually annotated sentiment corpus is available.

Table 9: Evaluation of Polarity Score of Developed Telugu SentiWordNet

Polarity	Positive	Negative
Percentage	82.0%	78.0%

8 Conclusion

Indian languages SentiWordNet(s) are being developed by **Dr Sentiment** involving Internet Population and this can be expanded by the use of dictionary based, WordNet based and corpus based approaches as described in (Das and Bandyopadhyay, 2010).

Our next target is to assign proper sense id to each SentiWordNet entry from AsianWordNet. Presently only the Bengali SentiWordNet[7] is downloadable from the author's web page.

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

References

Andreevskaia Alina and Bergler Sabine. CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In the Proceedings of the 4th SemEval-2007, Pages 117–120, Prague, June 2007.

Aue A. and Gamon M., Customizing sentiment classifiers to new domains: A case study. In the Proceedings of RANLP, 2005.

Das A. and Bandyopadhyay S. (2010). Phrase-level Polarity Identification for Bengali, In IJCLA, Vol. 1, No. 1-2, Jan-Dec 2010, ISSN 0976-0962, Pages 169-182.

Das A. and S. Bandyopadhyay. SentiWordNet for Indian Languages, In the 8th Workshop on Asian Language Resources (ALR), August 21-22, Beijing, China, 2010.

Esuli Andrea and Sebastiani Fabrizio. SentiWordNet: A publicly available lexical resource for opinion mining. In the Proceedings of LREC, 2006.

Mihalcea Rada, Banea Carmen and Wiebe Janyce. Learning multilingual subjective language via-cross-lingual projections. In the Proceedings of ACL, Pages 976–983, Prague, Czech Republic, June 2007.

Mohammad Saif, Dorr Bonnie, and Hirst Graeme. Computing Word-Pair Antonymy. In the Proceedings of EMNLP-2008, October 2008, Waikiki, Hawaii.

Pang Bo, Lee Lillian, and Vaithyanathan Shivakumar. Thumbs up? Sentiment classification using machine learning techniques. In the Proceedings of EMNLP, Pages 79–86, 2002.

Read Jonathon. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In the Proceedings of the ACL Student Research Workshop, 2005.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In Proceedings of LREC 2004, Pages 1083 – 1086, Lisbon, May 2004.

Voll, K. and Taboada, M. (2007). Not All Words are Created Equal: Extracting Semantic Orientation as a Function of Adjective Relevance. In the Proceedings of the 20th Australian Joint Conference on Artificial Intelligence. Pages. 337-346.

Wiebe Janyce and Mihalcea Rada. Word sense and subjectivity. In the Proceedings of COLING/ACL-06. Pages 1065-1072.

Wiebe Janyce and Riloff Ellen. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In the Proceeding CICLING, Pages 475–486, 2006.

Wilson Theresa, Wiebe Janyce and Hoffmann Paul (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In the Proceedings of HLT/EMNLP 2005, Vancouver, Canada.

Toward the Semantic Geospatial Web M. Egenhofer ACM-GIS 2002, McLean, VI A. Voisard and S.-C. Chen (eds.), Pages. 1-4, November 2002.

Susan Gaucha, Jason Chaffeeb and Alexander Pretschnerc. Ontology-based personalized search and browsing. In Proceeding of Web Intelligence and Agent Systems: An international journal. 2003. Pages 219–234. ISSN 1570-1263/03.

Gediminas Adomavicius and Alexander Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In the Proceeding of IEEE Transactions on Knowledge and Data Engineering, VOL. 17, NO. 6, June 2005. ISSN 1041-4347/05. Pages 734-749

Richard M. Tong. An operational system for detecting and tracking opinions in on-line discussion. In the Proceedings of the Workshop on Operational Text Classification (OTC), 2001.