Opinion-Polarity Identification in Bengali

Amitava Das* and Sivaji Bandyopadhyay**
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032, India
amitava.santu@gmail.com* and sivaji_cse_ju@yahoo.com**

Abstract— **In this paper, opinion polarity classification on news texts has been carried out for a less privileged language Bengali using Support Vector Machine (SVM)[1]. The present system identifies semantic orientation of an opinionated phrase as either positive or negative. The classification of text as either subjective or objective is clearly a precursor to determine the opinion orientation of evaluative text since objective text is not evaluative by definition. A subjectivity classifier has been used to perform sentence level subjectivity classification. The present system is a hybrid approach to the overall opinion polarity identification problem and works with lexicon entities and linguistic syntactic features. The baselines system works only with SentiWordNet (Bengali)[2]. The use of lexical features like negative words, stemming cluster, functional word and parts of speech improved the performance of the present system over baseline. Inclusion of the chunk feature has improved the precision of the system by 19.2%. A further improvement of 3.6% in precision of the system has been obtained with the use of dependency relations information. Evaluation results of the final system have demonstrated a precision of 70.04% and a recall of 63.02%.**

Keywords- Opinion Mining, Polarity Identification, Bengali and Phrase Level Polarity Identification.

I. INTRODUCTION

Sentiment recognition from text is a new subarea of Natural Language Processing (NLP) and has drawn considerable attention of the NLP researchers in recent times. Several subtasks can be identified within opinion mining; all of them involve tagging at document/sentence/phrase/word level according to expressed opinion. One such subtask is based on a given opinionated piece of text on one single issue or item, to classify the opinion as falling under one of two opposing sentiment polarities, or locate its position in the continuum between these two polarities. A large portion of work in sentiment-related classification/regression/ranking falls within this category. The binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion is called sentiment polarity classification or polarity classification. Much work on sentiment polarity classification has been conducted in the context of reviews (e.g., "thumbs up" or "thumbs down" for movie reviews) [1]. While in this context "positive" and "negative" opinions are often evaluative (e.g., "like" vs. "dislike"), there are other problems where the interpretation of "positive" and "negative" is subtly different. But development of a complete opinion mining system needs an automatic subjectivity detection module (it is a classification module that can differentiate among subjective or objective texts) followed by polarity classifier. The assumption that all texts are opinionated may make the system development easier but the resultant system will be unable to meet real life goal. The present system has been developed on news corpus which is more generic than review corpus. The baselines system works only with SentiWordNet (Bengali). The use of lexical features like negative words, stemming cluster, functional word, parts of speech improved the performance of the present system. Chunk feature has improved the precision of the system by 19.2%. A further improvement of 3.6% in precision of the system has been obtained with the use of information on the dependency relations. The system evaluation has shown the precision and recall values of 70.04% and 63.02% for Bengali respectively.

In this paper, a complete opinion mining system is described that can identify subjective sentences within a document and an efficient feature based automatic opinion polarity detection algorithm to identify polarity of phrases. Related works are described in Section II. Resource acquisition has been discussed in Section III. The feature extraction technique has been described in Section IV. Evaluation results have been presented in Section V and the conclusion has been drawn in Section VI.

II. RELATED WORKS

"What other people think" has always been an important piece of information for most of us in any decision-making process. An opinion could be defined as a private state that is not open to objective observation or verification [2]. 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 [3], subjectivity detection ([4], [5], [6], [7]), review mining [8], customer feedback [9] and strength of document orientation [10]. Methods on the extraction of opinionated sentences in a structured form can be found in [11]. Some machine learning text labeling algorithms like Conditional Random Field (CRF) [12], Support Vector Machine (SVM) [13] have been used to cluster same type of opinions. Application of machine-learning techniques to any NLP task needs a large amount of data. It is time-consuming and

[1] http://chasen.org/~taku/software/TinySVM/
[2] http://amitavadas.com/sentiwordnet.php

expensive to hand-label the large amounts 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 document 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. Some important works on opinion analysis in the newspaper domain are [14], [15] and [16], but no such efforts have been taken up in Indian languages especially in Bengali.

III. Resource Acquisition

The initiation of an opinion mining task for a new language demands sentiment lexicon and gold standard annotated data for machine learning and evaluation. The detail of resource acquisition process for annotated data, subjectivity classifier, sentiment lexicon and the dependency parser are mentioned below.

A. Corpus

Bengali is the fifth popular language in the World, second in India and the national language in Bangladesh. Automatic opinion mining or sentiment analysis task have mainly concentrated on English language till date. Bengali is a less computational privileged language. Hence Bengali corpus acquisition is an essential task for any NLP system development. For the present task Bengali news corpus has been identified. 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 have been identified as the relevant corpus in Bengali. A brief statistics about the corpus are reported in the Table I. The corpus is then manually annotated and used for training and testing respectively. Detailed reports about this news corpus development in Bengali can be found in [17].

TABLE I. Bengali News 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

B. Subjectivity Classifier

The subjectivity classifier as described in [17] has been used. The resources used by the classifier are sentiment lexicon, Theme clusters and POS tag labels.

The classifier first marks sentences that include opinionated words. In the next stage the classifier marks theme cluster specific phrases in each sentence. If any sentence includes opinionated words and theme phrases then the sentence is definitely considered as subjective. In the absence of theme words, sentences are searched for the presence of at least one strong subjective word or more than one weak subjective word for its consideration as a subjective sentence. The recall measure of the present classifier is greater than its precision value. The evaluation results of the classifier are 72.16% (Precision) on the NEWS Corpus.

The corpus is then validated by a human annotator and is effectively used during training and testing of the polarity classifier.

C. SentiWordNet (Bengali)

Words that are present in the SentiWordNet (Bengali) carry opinion information [18]. Sentiment lexicon is used as an important feature during the learning process. 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 supervised classifier. Words which are collected directly from SentiWordNet (Bengali) are tagged with positivity or negativity score. 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.

D. Dependency Parser

Dependency feature in opinion mining task has been first introduced in [12]. This 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. But unfortunately dependency parser for Bengali is not freely available. In this section we describe the development of a basic dependency parser for Bengali language.

The probabilistic sequence models, which allow integrating uncertainty over multiple, interdependent classifications and collectively determine the most likely global assignment, may be used in a parser. A standard model, Conditional Random Field (CRF)[3], has been used. The tag set that has been used here is same as used in the NLP Tool Contest in ICON 2009[4]. The input file in the Shakti Standard Format (SSF)[5] includes the POS tags, Chunk labels and morphology information. The statistical dependency parser for Bengali as described in [19] has been used in the present work.

IV. Features Extraction

SVM treats opinion polarity identification as a sequence tagging task. SVM views the problem as a pattern-matching

[3] http://crfpp.sourceforge.net
[4] http://ltrc.iiit.ac.in/icon2009/nlptools.php
[5] http://www.docstoc.com/docs/7232788/SSF-Shakti-Standard-Format-Guide

task, acquiring symbolic patterns that rely on both the syntax and lexical semantics of a phrase. We hypothesize that a combination of the two techniques would perform better than either one alone. With these properties in mind, we define the following features for each word in an input sentence. For pedagogical reasons, we may describe some of the features as being multi-valued (e.g. stemming cluster) or categorical (e.g. POS category) features. In practice, however, all features are binary for the SVM model. In order to identify features we started with parts of speech (POS) categories and continued the exploration with the other features like chunk, functional word, SentiWordNet (Bengali), stemming cluster, Negative word list and Dependency tree feature. The feature extraction pattern for any machine learning task is crucial since proper identification of the entire features directly affect the performance of the system. Functional word, SentiWordNet (Bengali) and Negative word list is fully dictionary based. On the other hand, POS, chunk, stemming cluster and dependency tree features are extractive. Classifying the polarity of opinionated texts either at the document/sentence or phrase level is difficult in many ways. A positive opinionated document on a particular object does not mean that the author has positive opinions on all aspects. Likewise, a negative opinionated document does not mean that the author dislikes everything. In a typical opinionated text, the author writes both positive and negative aspects of the object, although the general sentiment on the object may be positive or negative. Document-level and sentence-level classification does not provide such information. To obtain such details, there is a need to go to the object feature level.

A. *Parts Of Speech (POS)*

Number of research activities like [5] has proved that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion mining tasks, like the one presented in [20] are mostly based on adjective words.

B. *Chunk*

Chunk level information is effectively used as a feature in supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. It has been noted that 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' [21]. Similar fuzziness appeared in our marking of polar elements, such as 'কেন্দ্রীয় দলের দূর্নীতিতে' (corruption of central team) and 'দূর্নীতিতে' (corruption). Hence the hypothesis is to stick to chunk labels to avoid any further disambiguation. A detailed empirical study reveals that polarity clue may be defined in terms of chunk tags.

C. *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 syntactic pattern of an opinionated sentence. A list of 253 entries is collected from the Bengali corpus. First a unique high frequency word list is generated where the assumed threshold frequency is considered as 20. The list is manually corrected keeping in mind that a word should not carry any opinionated or sentiment feature.

iPhone is wonderful *and* easy to use.

In general writing practice positive words come together with conjunct like '*and*'. It is very hard to find 'wonderful and difficult to use' rather writers generally prefers '*but* 'instead of '*and*'.

D. *SentiWordNet (Bengali)*

Words that are present in the SentiWordNet (Bengali) carry opinion polarity information. Each word in a document is searched with its attached POS category in the SentiWordNet (Bengali). The obtained polarity class is used as a binary feature in the supervised classifier.

E. *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 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 same root form are grouped in a finite number of clusters with the identified root word as cluster center. For Bengali a stemming cluster based methodology has been used [22].

F. *Negative words*

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

G. *Dependency tree feature*

Dependency feature [21] has been successfully used here to identify modifier relationship of any polar phrase within a sentence. The analysis of Bengali corpora reveals that people generally use negation words/modifiers with any positive polar phrases. As an example

সে আদৌ ভালো নয়। (He is not good enough.)

The feature extractor module searches the dependency tree using breadth-first search to identify syntactically related nodes. The purpose of the feature is to encode dependency structure between related polar phrases.

V. EVALUATION

The evaluation result of the SVM-based polarity classification task for Bengali is presented in Table II. The evaluation result of the system for each polarity class i.e., positive and negative are mentioned separately in the Table III.

TABLE II. RESULTS OF POLARITY CLASSIFICATION.

Language	Domain	Precision	Recall
Bengali	***NEWS***	70.04%	63.02%

TABLE III. POLARITY WISE SYSTEM EVALUATION.

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

The effects of various features on the precision of the system have been studied. The phrase-level polarity identification task based on SentiWordNet has an accuracy of 47.60% and this is considered as the baseline. It may be observed from Table IV that incremental use of several lexical resources like SentiWordNet, 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 has been obtained. Further use of syntactic feature in terms of dependency relations has improved the system precision to 70.4% thus an increase of 3.6% in precision over the baseline has been obtained.

TABLE IV. FEATURE WISE SYSTEM PERFORMANCE.

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%

VI. CONCLUSION

One limitation of log-linear function models like SVM is that they cannot form a decision boundary from conjunctions of existing features, unless conjunctions are explicitly given as part of the feature vector. To maintain the granularity, features are explicitly mentioned as a classical word lattice model. A post-processor finally assigns the polarity value to the chunk head depending upon the chunk head resultant polarity.

REFERENCES

[1] Peter Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceeding of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics.

[2] Randolph Quirk, Sidney Greenbaum, Geoffry Leech and Jan Svartvik. A comprehensive Grammar of the English Language. Longman, New York. (1985)

[3] Tomohiro Fukuhara, Hiroshi Nakagawa and Toyoaki Nishida. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007.

[4] Baroni M and Vegnaduzzo S. Identifying subjective adjectives through web-based mutual information. Proceedings of Konvens, pages 17-24, 2004.

[5] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), 2000.

[6] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (ACL), pages 271-278, 2004.

[7] Soo-Min Kim and Eduard Hovy. Automatic detection of opinion bearing words and sentences. In Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2005.

[8] Hu and Liu. Mining and summarizing product reviews. Proceedings of 10th ACM SigKDD, 2004.

[9] Michael Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the International Conference on Computational Linguistics (COLING), 2004.

[10] Esuli A and Sebastini F. Determining the semantic orientation of terms through gloss analysis. Proceedings of CIKM, 2005.

[11] Nozomi Kobayashi, Kentaro Inui and Yuji Matsumoto. Opinion Mining from Web documents: Extraction and Structurization. Journal of Japanese society for artificial intelligence, Vol.22 No.2, special issue on data mining and statistical science, pages 227-238, 2007.

[12] Yejin Choi, Clarie Cardie, Ellen Riloff and Siddharth Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 355-362, 2005.

[13] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 412-418, 2004.

[14] S. Argamon-Engelson, M. Koppel, and G. Avneri. Style-based text categorization: What newspaper am I reading?. In Proceedings of the AAAI Workshop on Text Categorization, pages 1-4, 1998.

[15] L.-W. Ku, Y.-T. Liang, and H.-H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In Proceeding of AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 100-107, 2006.

[16] A. Stepinski and V. Mittal. A fact/opinion classifier for news articles. In Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR), pages 807-808, 2007.

[17] Amitava Das and Sivaji Bandyopadhyay, Subjectivity Detection in English and Bengali: A CRF-based Approach. In ICON 2009.

[18] A. Das and S. Bandyopadhyay. SentiWordNet for Bangla., In Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary , February 23th to 24th, 2010, Mysore.

[19] A. Ghosh, A. Das, P. Bhaskar, S. Bandyopadhyay. Dependency Parser for Bengali : the JU System at ICON 2009., In NLP Tool Contest ICON 2009, December 14th-17th, 2009, Hyderabad.

[20] Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 27-29, 2006.

[21] Wilson, Theresa and Wiebe, Janyce, Annotating Attributions and Private States, In Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky. pages 53--60, 2005.

[22] A. Das and S. Bandyopadhyay. Morphological Stemming Cluster Identification for Bangla., In Knowledge Sharing Event-1: Task 3: Morphological Analyzers and Generators, January 24th to 25th, 2010, Mysore.