Phrase-level Polarity Identification for Bangla

Amitava Das[1] and Sivaji Bandyopadhyay[2]

Department of Computer Science and Engineering
Jadavpur University, Kolkata 700032, India
amitava.santu@gmail.com[1] sivaji_cse_ju@yahoo.com[2]

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 determining the opinion orientation of evaluative text since objective text is not evaluative by definition. A rule based subjectivity classifier has been used. The present system is a hybrid approach to the problem, works with lexicon entities and linguistic syntactic feature. Evaluation results have demonstrated a precision of 70.04% and a recall of 63.02%.

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

1 Introduction

Emotion 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) [2]. 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

[1] http://chasen.org/~taku/software/TinySVM/

or objective texts) followed by polarity classifier. Assuming that all texts are opinionated may cause the system development easier but the resultant system will be unable to meet real life goal. Very little attempt could be found in literature to develop a complete opinion mining system. Rather people concentrate on specific sub problems. The present system has been developed on news corpus which is more generic than review corpus. The system evaluation has shown the precision and recall values are 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 2. Resource acquisition has been discussed in Section 3. The feature extraction technique has been described in Section 4. Conclusion has been drawn in Section 6.

2 Related Works

"What other people think" has always been an important piece of information for most of us during any decision-making process. An opinion could be defined as a private state that is not open to objective observation or verification [3]. 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 [4], subjectivity detection ([5], [6], [7],[8]), review mining [9], customer feedback [10] and strength of document orientation [11]. Methods on the extraction of opinionated sentences in a structured form can be found in [12]. Some machine learning text labeling algorithms like Conditional Random Field (CRF) ([13],[14]), Support Vector Machine (SVM) [15] 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 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 [16]. Some important works on opinion analysis in the newspaper domain are [17], [18] and [19], but no such efforts have been taken up in Indian languages especially in Bengali.

Various opinion mining methods have been reported that use lexical resources like WordNet [20], SentiWordNet [21] and ConceptNet [22] etc.

3 Resource Acquisition

To start 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.

3.1 Data

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 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 are identified as the relevant corpus in Bengali. A brief statistics about the corpus are reported in the Table 1. 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 [23]. The annotation scheme that has been used to annotate the corpus is reported in Table 2. The positive algebraic sign in the feature structure ("<fs af=,+,") depict the phrase polarity as positive.

Table 1. 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

Table 2. Bengali News Corpus Annotation Scheme

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

3.2 Subjectivity Classifier

The subjectivity classifier as described in [1] has been used. The resources used by the classifier are sentiment lexicon, Theme clusters and POS tag labels.

The classifier first marks sentences bearing 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.

3.3 Sentiment Lexicon

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 polarity: out of context, does the word seem to evoke something positive or something negative. For example, happy has a positive prior polarity, and sorrow has a negative prior polarity. However, the contextual polarity of a phrase in which a word appears may be different from the word's prior polarity. There are two main lexical resources widely used in English: SentiWordNet [21] and Subjectivity Word List [24] for Subjectivity Detection. SentiWordNet is an automatically constructed lexical resource for English which assigns a positivity score and a negativity score to each WordNet synset. Positivity and negativity orientation scores range within 0 to 1. Release 1.1 of SentiWordNet for English was obtained from the authors of the same. The subjectivity lexicon was compiled from manually developed resources augmented with entries learned from corpora. The entries in the subjectivity lexicon have been labeled for part of speech as well as either strong subjective or weak subjective depending on reliability of the subjective nature of the entry.

A word level translation process followed by error reduction technique has been used for generating the Bengali Subjectivity lexicon from English.

A subset of 8,427 opinionated words has been extracted from 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 words are discarded [24] from the Subjectivity word list as they are labeled as weakly subjective.

For the present task, a English-Bengali dictionary (approximately 102119 entries) developed using the Samsad Bengali-English dictionary[2] has been chosen. A word level lexical-transfer technique is applied to each entry of SentiWordNet and Subjectivity word list. Each dictionary search produces a set of Bengali words for a

[2] http://dsal.uchicago.edu/dictionaries/biswas_bengali/

particular English word. The set of Bengali words for an English word has been separated into multiple entries to keep the subsequent search process faster. The positive and negative opinion scores for the Bengali words are copied from their English equivalents. This process has resulted in 35,805 Bengali entries.

3.4 Dependency Parser

Dependency feature in opinion mining task has been first introduced by [25]. 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 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 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 identified from the training file and noted as an input feature in the CRF 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 CRF 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.

Root Word: Some dependency relations are difficult to identify without the word itself. It is better to come with some example.

AjakAla NN NP X k7t

In the previous example, there is no clue except the word itself. The word itself is noun, chunk level denotes a noun phrase and there is no vibhakti attached to the word. For these cases, word lists of temporal words, locations names and person names have been used for disambiguation [26]. Specifically identification of k7t relation is very tough because the word itself will be a common noun or a proper noun but the information of whether the word denotes a time or a location helps in the disambiguation.

[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

Part of Speech: Part of speech of a word always plays a crucial role to identify dependency relation. For example dependency relations like k1 and k2 in most of the cases involve a noun. It has been observed through experiments that not only POS tag of present word but POS tags of the context words (previous and next) are useful in identifying the dependency relation in which a word takes part.

Chunk label: Chunk label is the smallest accountable unit for detection of dependency relations and it is an important feature. But during the training sentences are parsed into word level, hence chunk label are associated to the appropriate words with the labels as B-X (beginning), I-X (Intermediate) and E-X (End) (where X is the chunk label).

Vibhakti: Indian languages are mostly non-configurational and highly inflectional. Grammatical functions (GFs) are predicted by case inflections (markers) on the head nouns of noun phrases (NPs) and postpositional particles in postpositional phrases (PPs). In the following example the '0_janya' vibhakti inflection of the word "pAoyZAra" leads to rh (Hetu - causal) case inflections. However, in many cases the mapping from case marker to GF is not one-to-one.

4 Features Extraction

SVM treats opinion polarity identification as a sequence tagging task. SVM views the problem as a pattern-matching 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 Part Of Speech (POS) categories and continued the exploration with the other features like chunk, functional word, SentiWordNet in Bengali[1], 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 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.

4.1 Part Of Speech (POS)

Number of research activities like [6], [27] etc. have proved that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion mining tasks, like the one presented in [28], are mostly based on adjective words.

4.2 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' (Wilson and Wiebe, 2003). 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.

4.3 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 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.

4.4 SentiWordNet

Words that are present in the SentiWordNet carry opinion information. The developed 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 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 SentiWordNet are tagged with positivity or negativity score.

4.5 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 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. Details could be found in [30].

4.6 Negative words

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

4.7 Dependency tree feature

Dependency feature 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.

5 Evaluation

The evaluation result of the SVM-based polarity classification task for Bengali is presented in Table 3. The evaluation result of the system for each polarity class i.e., positive and negative are mentioned separately in the table 4.

Table 3. Results of Polarity classification.

Language	Domain	Precision	Recall	
Bengali	NEWS	70.04%	63.02%	

Table 4. Polarity wise System Evaluation.

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

6 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's resultant polarity. We are now working on improving the performance of the present system. Future task will be in the direction of development techniques for creation of opinion summaries according to their polarity classes.

References

1. Amitava Das and Sivaji Bandyopadhyay. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009).
2. 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.
3. Randolph Quirk, Sidney Greenbaum, Geoffry Leech and Jan Svartvik. A comprehensive Grammar of the English Language. Longman, New York. (1985)
4. 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.
5. Baroni M and Vegnaduzzo S. Identifying subjective adjectives through web-based mutual information. Proceedings of Konvens, pages 17-24, 2004.
6. 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.
7. 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.
8. 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.
9. Hu and Liu. Mining and summarizing product reviews. Proceedings of 10th ACM SigKDD, 2004.
10. 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.
11. Esuli A and Sebastini F. Determining the semantic orientation of terms through gloss analysis. Proceedings of CIKM, 2005.
12. 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.
13. 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.

14.Andrew Smith, Trevor Cohn and Miles Osborne. Logarithmic Opinion Pools for Conditional Random Fields. In Proceeding of the 43rd Annual Meeting of the ACL, pages 18-25, 2005.
15.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.
16.Gregory Grefenstette, Yan Qu, James G. Shanahan and David A. Evans. Recherché d'Information Assistée par Ordinateur. In Proceedings of RIAO, 7th International Conference on 2004.
17. 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.
18. 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.
19. 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.
20. A. Esuli and F. Sebastiani. Page Ranking WordNet synsets: An application to opinion mining. In Proceedings of the Association for Computational Linguistics (ACL), 2007.
21. Esuli and Sebastini. Sentiwordnet: a publicly available resource for opinion Genova, Italy. 2006.
22. Nathan Eagle, Push Singh and Alex (Sandy) Pentland .Common sense conversations: understanding casual conversation using a common sense database. In Proceedings of the Artificial Intelligence, Information Access, and Mobile Computing Workshop (IJCAI 2003).
23.Ekbal, A., Bandyopadhyay. S. A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation Journal. pages 173-182, 2008.
24. Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In Proceedings of CICLing 2005 (invited paper).
26.Choi, Yejin and Cardie, Claire and Riloff, Ellen and Patwardhan, Siddharth, Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing. Pages 355-362, 2005.
27.Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay. 2008. Language Independent Named Entity Recognition in Indian Languages. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages. pages 33-40.
28.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.
29.Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the Conference on Knowledge Capture (K-CAP), pages 70-77, 2003.
30. Amitava Das and Sivaji Bandyopadhyay, Subjectivity Detection in English and Bengali: A CRF-based Approach. In ICON 2009.