Subjectivity Detection in English and Bengali: A CRF-based Approach

Amitava Das
Department of Computer Science and Engineering
Jadavpur University
Jadavpur University, Kolkata 700032, India
amitava.santu@gmail.com

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
Jadavpur University, Kolkata 700032, India
sivaji_cse_ju@yahoo.com

Abstract

With the proliferation of online reviews and sentiments the Web is becoming more and more useful and important information resource for people. As a result, automatic opinion/sentiment mining has become a hot research topic recently. Extracting opinions from text is a hard semantic problem. Subjectivity Detection is studied as a text classification problem that classifies texts as either subjective or objective. This paper illustrates a Conditional Random Field (CRF) based Subjectivity Detection approach tested on English and Bengali multiple domain corpus to establish its effectiveness over multiple domain perspective. The motivation is to develop generic domain independent solution architecture for a less computerized language like Bengali. A relatively simple and less human interactive technique has been proposed for developing opinion mining resources for Bengali. The features used in the CRF-based classifier could be extracted for any new language with minimum linguistics knowledge. The final classifier has resulted precision values of 76.08% and 79.90% for English and 72.16% and 74.6% for Bengali for the news and blog domains respectively.

1 Introduction

Opinion and sentiment are individual's intuitions expressed diversely in text. Extracting the emotive content of text, rather than factual content, is a complex problem and is known as a Subjectivity Detection. Most of the research on Subjectivity detection to date has focused on English, mainly due to the availability of resources for subjectivity analysis, such as sentiment lexicons and manually labeled corpora. But no such effort has been noticed yet for Indian languages including Bengali. The research contribution of this work lies in the identification of most sophisticated and inexpensive way to automatically develop sentiment lexicon for a new language like Bengali and the identification of various features for the Subjectivity Classification algorithms. The features used need limited linguistic knowledge and thus can be easily ported to a new language. The proposed CRF-based Subjectivity Classification algorithm has been compared with standard techniques applied for English and favorable results have been obtained.

2 Related Works

Development of Subjectivity Classifier for a new language demands sentiment lexicon and gold standard annotated data for machine learning and evaluation.

Rada 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 the present work, the same translation based technique has been followed but with some important modification for Subjectivity lexicon generation for Bengali. A stemming cluster technique followed by SentiWordNet validation

Proceedings of ICON-2009: 7th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009

has been proposed. This technique helps to translate ambiguous Subjective words that may lose their subjective meaning once lemmatized. The details of Sentiment Lexicon generation are mentioned in Section 3.

In case of sentence level subjectivity annotation a parallel corpus based approach has been proposed in Rada Mihalcea et al. (2007). But for Indian languages especially Bengali it is very hard to collect appropriate parallel corpora. Xiaojun Wan (2008) has proposed generation of Chinese reviews from English texts by Machine Translation. Publicly available tools like GoogleTrans, Yahoo Babel Fish and a word level translation module have been used. There is no publicly available Machine Translation system involving Bengali though an English-Hindi machine translation system is available in GoogleTrans. In the present work, a rule based sentence level subjectivity annotation has been done and finally human annotator checks it for validation.

Another significant effort on subjectivity annotation is found in Anthony et al. (2005). Opinion/Sentiment mining is identified as a very domain specific problem (in the present work it is described as language dependent). 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 new domain. The aim of the present paper 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., Namrata Godbole et al. (2007).

3 Sentiment Lexicon Generation

There are two main lexical resources widely used in English: SentiWordNet (Esuli et. al., 2006) and Subjectivity Word List (Wiebe and Riloff, 2005) 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. The essential issue in the present task is to select either the SentiWordNet or Subjectivity Word List as the best source lexical resource. A detailed analysis of the two lexical resources revealed some special characteristics as specified in the following Table 1.

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

Table 1. Statistics of both resources

It has been observed that 64% of the single word entries are common in the Subjectivity Lexicon and SentiWordNet. Instead of taking any one of the English lexical resources, it has been decided to generate a merged sentiment lexicon from both the resources by removing the duplicates. The new list consists of 14,135 numbers of tokens. Several filtering techniques have been used to generate the new list.

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 (as in Wiebe and Riloff, 2005) from the Subjectivity word list as they are labeled as weakly subjective.

In the next stage the words whose POS category in the Subjectivity word list is undefined and tagged as "*anypos*" are considered. These words may generate sense ambiguity issues in next stages of subjectivity detection. The words are checked in the SentiWordNet list for validation. If a match is found with certain POS category, the word is added to the new subjectivity word list. Otherwise the word is discarded to avoid ambiguities later.

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.

For the present task, a English-Bengali dictionary (approximately 102119 entries) developed using the Samsad[1] Bengali-English dictionary 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 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.

4 Language-Domain Dependent Knowledge Gathering

Sentiment in different domains can be expressed in very different ways (Engstr¨om, 2004). In the following two example sentences, the word 'heavily' appears in both.

a. It is *heavily* raining.
b. Into the issue of Nandigram and Lalgarh Governor *heavily* reacted.

In the second sentence, the word 'heavily' carries an opinion orientation. The system requires domain dependent knowledge to identify such situation. In the example sentence (b) the domain dependent clue tokens are "Nandigram and Lalgarh" and "Governor".

Therefore a Theme cluster detection technique has been evolved to capture the domain dependent knowledge. The technique is fully rule-based and works in two stages.

In the first stage, a document wise high frequent n-gram word list has been prepared. Only four POS categories (Noun, Verb, Adjective and Adverb) are considered for this stage as they are more opinionated (Hatzivassiloglou et. al., 2000). The Sanford POS tagger has been used for English POS identification task and a CRF[2] based POS and chunking engine has been developed for Bengali.

In the second stage, documents are clustered according to their theme. If 30% of theme expressions of document D_A matches with document D_B then the two documents are assigned to the same cluster assuming that they are on related topic. Starting from N clusters (where N is the total number of documents in the domain specific corpus) the system iterates several times and finally assigns documents into a finite number of K (Where $K<N$) clusters. Every theme is then identified by the total number of theme expressions of the documents into the particular theme cluster.

5 Data

Manually annotated Subjective data is available for English in the form of Multi Perspective Question Answering (MPQA)[3] corpus and International Movie Database (IMDB)[4] among others. Such manually developed data is not available for Indian languages and Bengali is not an exception. In the present work, a rule based sentence level subjectivity annotation has been done for Bengali that is finally checked for validation by human annotator. The complete manual development of the annotated corpus would be expensive. But the present technique is relatively simple and less human interactive that can be followed for any new language with limited number of resources. A simple interface has been designed for human validation

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

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

[3] http://www.cs.pitt.edu/mpqa/databaserelease/

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

of annotated data. The tool highlights the words present in sentiment lexicon by four different colors within a document according to their POS categories (Noun, Adjective, Adverb and Verb). Some statistics about the Bengali news corpus and blog corpus on which the annotation has been carried out are represented in the Table 2. The MPQA and the IMDB corpus have been selected for English.

	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 Bengali Corpus Statistics

5.1 Rule based Subjectivity Classifier

The subjectivity classifier as described in (Das and Bandyopadhyay, 2009) 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. Every opinionated word is validated along with its POS tag in the developed sentiment lexicon (with appropriate POS tags). Words for which the POS category obtained from the POS tagger does not match in the sentiment lexicon are discarded.

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 reported in (Das and Bandyopadhyay, 2009).

Both the news and blog corpus is then validated by a human annotator and is effectively used in a supervised subjectivity classifier for training and testing.

6 Supervised Subjectivity Classifier

Each document is represented as a feature vector for machine learning task. After a series of experiments the following feature set is found to be performing well as a subjectivity clue.

6.1 Part of Speech

Number of research activities like Hatzivassiloglou et. al. (2000), Chesley et. al. (2006) 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 Nasukawa et. al. (2003), are mostly based on adjective words.

6.2 Chunk

Chunk label 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. In the task of identification Theme expressions chunk label markers play a crucial role. A detailed empirical study reveals that Subjectivity clue may be defined in terms of chunk tags. Details could be found in the Das and Bandyopadhyay (2009).

6.3 Sentiment Lexicon

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 strong or weak subjective strength measure. Such 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. 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 score respectively.

6.3.1 Stemming

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. Examples have been shown in Table 3.

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 3. POS Category wise SpellingVariations.

6.4 Frequency

Frequency always plays a crucial role in identifying the importance of a word in the document. After Function word removal and POS annotation, system generates four separate high frequent word lists for each of the four POS categories: Adjective, Adverb, Verb and Noun. Word frequency values are then effectively used as a crucial feature in the Subjectivity classifier.

6.5 Positional Aspect

Depending upon the position of subjectivity clue, every document is divided into a number of zones, such as Title of the document, the first paragraph and the last two sentences. A detailed study was done on the MPQA and Bengali corpus to identify the roles of the positional aspect in the detection of subjectivity of a sentence and these results are shown in Table 4. Zone wise statistics could not be done for the IMDB corpus because the corpus is not presented as a document.

6.5.1 Title of the document

It has been observed that the Title of a document always carries some meaningful subjective information. Thus a Thematic expression bearing title words (words that are present in the title of the document) always get higher score as well as the sentences that contain those words.

6.5.2 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 bearing Thematic Expressions.

6.5.3 Last Two Sentences

It is a general practice of writing style that every document concludes with a summary of the opinions expressed in the document.

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

Table 4 Statistics on Positional Aspect.

7 Average Distribution

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 observed encouraging results to identify the theme of a document. These distribution rules are identified after analyzing the English corpora and the same rules are applied to Bengali. An increment of 0.83% and 0.65% has been found respectively for theme detection in English and Bengali corpora after application of the distribution rules.

8 Evaluation Result

The main motivation of the present work is to establish how the proposed method can be adopted for a new language.

The effectiveness of each feature in the CRF-based Subjectivity classification task for English and Bengali are presented in Table 5. The precision and recall values are shown in Table 6 for all

the corpora selected for English and Bengali. It can be 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 theme clusters.

Features	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 5. Feature wise System Performance

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 6. Overall System Performance

9 Conclusion

In this paper the task of binary subjectivity classification in Bengali has been studied. Furthermore, a general architecture for developing subjectivity resources for any new language has been presented. According to the study, emotion and topic classification of Bengali articles can be improved by using a classification approach achieved by Theme cluster detection technique. Theme cluster detection technique improves the subjectivity classification task by finding out the importance of the words other than opinion/sentiment bearing words. The future plan is to extend the work in the direction of developing a relational network of Theme words and opinion/sentiment words. A relational network structure will illustrate the semantic relations between opinionated and non-opinionated entities in any document.

References

A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," In the Proceedings of Recent Advances in Natural Language Processing (RANLP), 2005.

Amitava Das and Sivaji Bandyopadhyay. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009) (Accepted).

Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of Language Resources and Evaluation (LREC), 2006.

E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In Proceedings of CICLing 2005 (invited paper).

N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," In Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007.

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.

Rada Mihalcea, Carmen Banea and Janyce Wiebe. Learning Multilingual Subjective Language via Cross-Lingual Projections. In Proceeding of Association for Computational Linguistics. Pages 976—983. 2007.

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.

Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), pages 299-305, 2000.

Xiaojun Wan. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In Proceeding of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 553–561, 2008.