Theme Detection an Exploration of Opinion Subjectivity

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

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

Abstract

Work in opinion mining and classification often assumes the incoming documents to be opinionated. Opinion mining system makes false hits while attempting to compute polarity values for non-subjective or factual sentences or documents. It becomes imperative to decide whether a given document contains subjective information or not as well as to identify which portions of the document are subjective or factual. In this work a Theme Detection technique has been evolved for more generic domain independent subjectivity detection that classifies sentences with binary feature: opinionated or non-opinionated. Theme Detection technique examines sentence level opinion and finally accumulates the opinion clues to reach the discourse level subjectivity. The subjectivity detection system has been evaluated on the Multi Perspective Question Answering (MPQA) corpus as well as on Bengali corpus. The system evaluation has shown the precision and recall values of 76.08 and 83.33 for English and 72.16 and 76.00 for Bengali respectively.

Keywords: Opinion Mining, Subjectivity Detection, Theme Detection.

1. Related Works

An opinion could be defined as a private state that is not open to objective observation or verification [7]. 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 [8], subjectivity detection ([9],[10],[11],[12]), Methods on the extraction of opinionated sentences in a structured form can be found in [6]. Some machine learning text labeling algorithms like Conditional Random Field (CRF) ([16],[17]), Support Vector Machine (SVM) [18] 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 [19]. Some important works on opinion analysis in the newspaper domain are [20], [21] and [22], but no such efforts have been taken up in Indian languages especially in Bengali.

2. Annotated Data Preparation

The opinion subjectivity detection technique presented in this paper is rule based in nature and hence annotated data preparation is necessary for system testing and evaluation. The technique has been applied on both English and Bengali language texts. In case of English we choose MPQA corpus [27]. Since in the MPQA corpus the private states in a sentence are annotated, sentence level opinion subjectivity annotation has been done as described in Section 3.1. No such corpus is available for Bengali. Hence, the editorial pages, i.e., 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. Detailed reports about this news corpus development in Bengali can be found in [26]. The relevant Bengali news corpus is hand annotated as described in Section 3.2.

2.1. Subjectivity-Identification in MPQA

The annotation scheme in the MPQA corpus identifies key components and properties of opinions, emotions, sentiments, speculations, evaluations, and other private states [27]. The properties of a private state frame include the source of the private state (i.e., whose private state is being expressed), the target (i.e., what the private state is about), and various other details involving intensity, significance and type of attitude. The annotation within the MPQA corpus is not at sentence level but at word or phrase level. Private states in MPQA are classified into two basic categories i.e.

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

direct subjective frames and expressive subjective element. To distinguish opinion-oriented materials from other factual materials, objective speech event frames have been defined.

Direct Subjective Frame: A private state containing direct subjective element is called a direct subjective frame. For example, in the sentence

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

The word 'fears' represents a private state and is annotated as 'Direct Subjective Frame'.

Expressive subjective elements: A private state containing no direct opinion but only subjective references to opinion is called an expressive subjective element frame. For example, in the sentence

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

The phrase "full of absurdities" represents a private state and is annotated as 'Expressive Subjective Element'.

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

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

During the subjective tagging of MPQA sentences, the hypothesis is that if a sentence has any Direct or Expressive Subjective phrases then the sentence would be subjective itself. Hence, the sentences containing either of the two private states (i.e. Direct Subjective and Expressive Subjective) are extracted as subjective sentences and the sentences containing only objective speech event or no annotated private states are discarded. There are features of an attitude frame like intensity in the MPQA annotation scheme. But for the present task other attributes of an attitude frame are not considered.

2.2. Subjectivity-Annotation in Bengali

From the collected document set (Letters to the Editor Section), some documents have been chosen for the annotation task. Documents that appeared within an interval of four months are chosen on the hypothesis that these letters to the editors will be on related events.

A simple annotation tool has been designed for annotating the subjective sentences. Three annotators participated in the present task. The documents with such annotated sentences are saved in XML format.

The tool also highlights the sentiment words (SentiWordNet, see section 4.2.5 for details) by four different colors within a document according to their POS categories (Noun, Adjective, Adverb and Verb). This technique helps to increase the speed of annotation process. Finally 100 annotated documents have been produced. No inter-annotator agreement has been calculated for the present task. Some statistics about the Bengali news corpus is represented in the Table 1.

Total number of documents in the corpus	100
Total number of sentences in the corpus	2234
Avgerage number of sentences in a document	22
Total number of wordforms in the corpus	28807
Avgerage number of wordforms in a document	288
Total number of distinct wordforms in the corpus	17176

Table 1. Bengali News Corpus Statistics

3. Theme Detection

Theme Detection is a rule based algorithm to identify subjective sentences in text documents. The algorithm takes the subjectivity decision on the basis of several features of the sentences in the text. These features are obtained using various machine learning algorithms. Various linguistic resources that are used to derive several binary features have been developed manually. Theme detection process works in two stages. Theme detection technique first captures discourse level opinion theme in terms of thematic expressions which best describes the opinionated theme of a document. In the next level the algorithm examines the presence of thematic expression as an opinion constituent (Subject-Aspect-Evaluation) in any sentence. Subjectivity detection by lexicons like SentiWordNet or Subjectivity word list has been explored by researchers. Theme detection technique works on discourse level and takes care of syntactic structure of sentences. . The challenge is to identify the most concise feature set and construct the rules effectively for the two stage identification problem. Experiments are carried out with an initial list of features and finally some of the features are discarded as they are found to have no contribution towards increasing system performance. The Theme detection technique has been applied on both English and Bengali language texts. Motivation for the technique has been presented in Section 4.1. The various subjectivity clues or features and how these can be obtained have been discussed in Section 4.2. The evaluation results presented in the Section 6 show the effectiveness of the algorithm.

3.1. Motivation

Many supervised and unsupervised techniques have been explored for subjectivity annotation task by many researchers over a long period of time. Several linguistic resources and tools like dependency parsing, Named Entity Recognition, Morphological Analyzer, Stemmer, SentiWordNet, WordNet etc have been used several times in the subjectivity detection task. But in the case of morphologically rich Indian languages like Bengali, such resources and tools are not readily available. Highly inspired by Janyce Wiebe et.al, 2005 [28] the present work is initiated to develop a subjectivity classifier that will work on un-annotated text. Our aim is to design an automatic process that learns linguistically rich extraction patterns for subjective (opinionated) expressions and produces a rich ontological languagespecific (rather than domain dependent) knowledge. Subjective remarks come in a variety of forms, including opinions, rants, allegations, accusations, suspicions, humor and speculations.

3.2. Learning Subjective Clues

Existing methods for opinion extraction tend to rely on relatively simple proximity-based or pattern-based techniques. However, these pattern-based techniques are not enough to extract opinions because these patterns can apply to the cases where all constituents of opinion appear in a sentence. It has been observed that most of the opinion constituents do not have a direct syntactic dependency relation within a sentence, mostly due to elliptical arguments. Based on a corpus study, it is proposed to define an opinion unit as a quadruple, i.e., the opinion holder, the subject being evaluated (Subject), the part or the attribute in which it is evaluated (Aspect), and the evaluation that expresses a positive or negative assessment (Evaluation). The present subjectivity detection algorithm has been applied to News corpus (both for English and Bengali) where name of the author of any article is rarely mentioned. Hence, the Opinion Holder information is not taken into consideration and only the Subject-Aspect-Evaluation constituents are identified for analysis. These constituents can be further defined as:

Subject: A named entity (Person or Location etc.) of a given particular class of interest (e.g. a leader name or a location name where an incident occurred).

Aspect: An attribute of the subject with respect to which evaluation is made (size, color, date etc.). The aspect can define a characteristic of the subject or an integral part of the subject.

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

Initially, a detailed analysis of the English MPQA and Bengali newspaper corpus has been done to understand the most concise and effective features for opinion (i.e., opinion constituents) identification and their characteristics in the corpus. 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, ontology list, SentiWordNet, stemming cluster, frequency, positional aspect (e.g. title, first Paragraph, last two sentences, critical Issues) and average distribution. Each of the features and the methods for their identification are now being discussed.

3.2.1 Part Of Speech (POS)

Hatzivassiloglou et. al., 2000 [29], Chesley et. al., 2006 [30] etc. have proved that those words carrying opinion in sentences are mainly adjective, adverb, noun and verbs. Many opinion mining task, for example the one presented in [31], are mostly based on adjective words. This means that this part-of-speech (POS) tag is more important for Subjectivity Detection than others. The identified POS categories that carry opinion information are Adjective, Adverb, Verb and Noun while the opinion information for words of other POS categories is difficult to generalize. Thus, we concentrated on identifying the POS categories of the words, especially to see whether these words are adjectives, adverbs, noun and verbs.

Stanford Parser[1] has been used for English text to get the word level POS category. The overall accuracy of this tool as reported in Klein et.al. 2003 [32] is 86.7%.

The POS Tagging Engine for Bengali text has been developed using the statistical Conditional Random Fields (CRF)[2] [33]. The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various part of speech (POS) classes. The training set consists of 200K words and has been manually annotated with a POS tag-set[3] of twenty one tags developed by International Institute of Information Technology Hyderabad (IIIT-H). The system produces output in Shakti-Standard-Format[4] (SSF), also developed by IIIT-H for Indian languages. Experimental test results show the effectiveness of the CRF based POS tagging system with an overall average 87.23% accuracy. Feature selection plays a crucial role in CRF framework. Experiments were carried out to find out the most suitable features for POS in Bengali. The Experimental results are shown below in Table 2.

Training-Set	Test-Set	Accuracy
16397	4587	87.23%

Table 2. Experimental Result of POS Tagging.

3.2.2 Chunk

Identification of subjective feature depends on opinion constituents. A detailed empirical study reveals that Subject, Aspect or Evaluation expressions may be defined in terms of chunk tags.

- **Subject phrases** are generally noun phrases with noun-noun, adjective-noun or noun-other combinations. The nouns in the subject phrases are generally named entities or low frequency noun words or out of vocabulary words. In case of multiple noun phrases in a sentence the head noun phrase is treated as a Subject phrase. Some empirical rules have been defined to identify the head noun phrase in a sentence: the distance of each NP (Noun phrase) from VP (Verb Phrase) is calculated in terms of characters. The NP which is situated at the farthest distance from VP is selected as the head NP. In case of multiple VPs, only the phrases tagged as Verb Finite Phrases are considered.

[1] http://nlp.stanford.edu/software/lex-parser.shtml#Download
[2] http://crfpp.sourceforge.net
[3] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf
[4] http://ltrc.iiit.net/MachineTrans/research/tb/shakti-analy-ssf.pdf

- **Aspect phrases** are an attribute of the Subject. But in News domain aspect is not only sub part of the Subject but also a conceptual sub-event or an actor's (i.e. Subject) activity on a particular event. More specifically the aspects are factual information which should be discarded by subjectivity detector. As an example:

1) Saddam hanged up yesterday night.

2) USA president says "He deserved it".

3) USA people's general opinion this is a form of real cruelty.

In the sentence 2, the aspect is Saddam Hussain's death is missing, which is an event reported in the sentence 1. In the sentence 3 subject is peoples of USA but the aspect is also an event that is president's statement which is a subsequent event of the main event i.e., Saddam Hussain's death is missing. Aspects are either missing or present as elliptical arguments in News corpus. In other cases when aspect present in a sentence, it has been observed that it is generally a noun phrase with some sentiment words. A small set of rules are defined for aspect identification in a sentence.

- **Evaluation phrases** are used to express an evaluation or the opinion holder's mental/emotional attitude. Detailed analysis of both the English and Bengali corpus has shown that Evaluation phrases are generally verb phrases with some evaluative sentiment phrases (adjective or adverbial). In case of missing finite verb phrases (Sentence 3 in the preceding example) in any sentence the evaluative phrases are generally adjective or adverbial phrases or noun phrases along with evaluative sentiment words (obtained from SentiWordNet). Rules have been defined accordingly to identify those phrases.

The common feature among the three constituents of opinion is that they could be noun phrases. Positional information and language features have been used to disambiguate among the three constituents.

The Stanford Parser was used for English chunking task. The parser produced a parse tree as the output that is converted into SSF format for further processing.

The chunker for Bengali texts is trained on the feature templates for predicting the chunk boundary tags using CRF. The accuracy of the chunker is shown in Table 3.

Training-Set	Test-Set	Accuracy
16397	4587	79.51%

Table 3. Experimental Result of Chunker.

3.2.3 *Functional Word*

Function words in a language are high frequency words and these words generally do not contribute to identify subjectivity; hence these words are dropped by system in the first stage. But function words help many times to understand syntactic pattern of an opinionated sentence, hence some rules are constructed based on functional words at the POS and Chunk level instead of the word itself. An example may illustrate the situation.

4) President and Army chief both congratulated NASA on the success of their expedition.

In the sentence 4 the subject is "President and Army chief" and the clue is the function word "and". Rules are defined to find out more than one consecutive NPs (Noun phrase) connected with any conjunct.

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. For English the functional word list is collected from Website[5]. In the English functional word list there are 300 entries.

3.2.4 *Ontology List*

Four different ontology lists corresponding to the four POS categories (Noun, Verb, Adjective and Adverb) of words have been prepared by selecting the top 100 words in each category from POS tagged English and Bengali texts. During the second stage of subjectivity detection, i.e., Theme Sentence Identification, words with Noun POS categories were separated into Named Entities and Common nouns.

3.2.5 *SentiWordNet in Bengali*

Words that are present in the SentiWordNet carry opinion information. SentiWordNet [34] is an automatically constructed lexical resource for English which assigns a positivity score, a negativity score, and a neutrality score to each WordNet synset. Release 1.1 of SentiWordNet for English was obtained from the authors. SentiWordNet Release 1.1 consists of 115,341 words marked with positive and negative orientation scores ranging from 0 to 1. A subset of 8,427 opinionated words was extracted from SentiWordNet, by selecting those whose orientation strength is above a threshold of 0.4.

As there was no such SentiWordNet for Bengali, a task has been initiated to develop a similar resource for Bengali. For the task, Samsad[6], a widely used English-Bengali dictionary available both in offline and online version, is selected. The Samsad English–Bengali dictionary has approximately 102119 entries. A word to word simple lexical-transfer technique is applied to each entry of SentiWordNet. Each dictionary search produces a set of Bengali words for a particular English word. Instead of making them into one entry we separate them into multiple one word entries for making 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 20,789 Bengali entries which is a useful resource in the Bengali Opinion Mining task.

The words in the POS tagged English and Bengali texts with the following POS tags, namely Noun, Verb Adjective and Adverbs, were checked into the

[5] http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

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

SentiWordNet for the respective languages. Words for which the POS category obtained from the POS tagger does not match in the SentiWordNet are discarded. Other words are considered important for subjectivity detection.

3.2.6 *Stemming Cluster*

Several words in a sentence that carry opinion information may be present in a sentence in their 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 developed. This feature 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. The term prefix/suffix is a sequence of first/last few characters of a word, which may not be linguistically meaningful. The use of prefix/suffix information works well for highly inflected languages like the Indian languages. Experiments are carried out with two types of algorithms: simple suffix stripping algorithm and score base stemming cluster identification algorithm. A small list of 205 suffixes for Bengali has been manually generated. The Suffix stripping algorithm simply checks if any word has any suffixes (one or more than one suffixes) from the list and then the word is assigned to the appropriate cluster where cluster center is the assumed root word, i.e., the form obtained after deleting the suffix from the surface form. Suffix stripping algorithm works well for Noun, Adjective, Adverb categories. In case of Verbs in Bengali, root form of the word changes when suffixes are added. Hence for the Bengali Verb words simple suffix stripping does not work well. The score based stemming technique has been designed to resolve the stem for inflected Verb words. The technique uses Minimum Edit Distance method [35], well known for spelling error detection, to measure the cost of classifying every word being in a particular class. Score based technique considers two standard operations of Minimum Edit Distance, i.e., insertion and deletion. The consideration range of insertion and deletion for the present task is maximum three characters. The idea is that the present word matches an existing cluster centre after insertion and/or deletion of maximum three characters. The present word will be assigned to the cluster that can be reached with minimum number of insertion and/or deletion. This is an iterative clustering mechanism for assigning each word into a cluster. The system iterates 6 times i.e. it starts from -3 (deletion of three characters) and ended with +3 (insertion of three characters) value and finally generate a finite number of stemming clusters. A separate list of verb inflections (only 50 entries) has been maintained to validate the result of the score based technique. The standard K-means Clustering technique has been used here. Each cluster center is treated as a root stem. For English, standard Porter Stemmer[7] algorithm has been used.

3.2.7 *Frequency*

Frequency 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 four POS cate-gories: Adjective, Adverb, Verb and Noun. The Theme Expression identification module then starts to recognize most important Theme Expressions from the four categorical lists. These lists contain single word and multiword entities (identified through chunk level information) simultaneously. The system makes several iterations to calculate the presence of each Theme Expressions in the opinion constituents (Subject-Aspect-Evaluation) in a sentence and update the associated score of the theme sentence. The system then proceeds to examine the valence of every sentence based on the presence of Thematic Expressions along with many other features and rule set.

Type	Root	Surface Form	Suffixes
Noun	ভারত	ভারতে, ভারতের	ে , ের
Adjective	অমানব, দুর্ভাগ্য	অমানবিক, দুর্ভাগ্যবশত	িক বশত
Adverb	ভারী, দূর, দূর	ভারিক্কি, দূরীভূত, দূরীকৃত	িক্কি, ীভূত, ীকৃত
Verb	খা	খাচ্ছেন, খেয়েছিলেন	চ্ছেন, য়েছিলেন

Table 4. POS Category wise Variations.

3.2.8 *Positional Aspect*

Depending upon the position of subjectivity clue, every document is divided into a number of zones. The dependency factors of this feature are Title of the document, first paragraph and 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 Tables 5.

3.2.8.1 *Title of the document*

It has been observed that Title of a document always carries some meaningful subjective information. Thus a Thematic expression containing title words (words that are present in the title of the document) always get higher score. The sentences that contain these thematic expressions also get higher scores.

3.2.8.2 *First Paragraph*

It has been observed that people generally give a brief idea of their beliefs and speculations in the first paragraph of the document and subsequently elaborate

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

or support their ideas with relevant reasoning or factual information. This first paragraph information is useful in both stages of Thematic Expressions detection and the detection of subjective sentences bearing Thematic Expressions. The necessary set of rules is constructed accordingly.

3.2.8.3 *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
First Paragraph	48.00%
Last Two Sentences	64.00%

Table 5 Statistics on MPQA

Positional Factors	Percentage
First Paragraph	56.80%
Last Two Sentences	78.00%

Table 6 Statistics on Bengali Corpus

3.2.9 *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 values. Thematic words that are well distributed throughout the document are important thematic words. A threshold range of a distance of 10-17 words between any two occurrences of a thematic word is used to identify well distributed thematic words. An increment of 0.83% and 0.65% found respectively for theme detection in English and Bengali corpora have been observed after application of the distribution rules.

4. Experimental Results

The Baseline systems for both the English and Bengali languages are developed using the rules that are based on two primary features i.e. frequency and positional information. The result of the Baseline systems is reported in the Table 7.

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

Table 7. Results on Base-Line System

Further incremental improvement of the baseline system depends on selection of appropriate features for inclusion. The features as defined in Section 4.2 are considered for experimentation. Some of the features are discarded as they are found to have no contribution towards increasing the system performance. The final list of features for which any incremental improvement towards system performance is observed is reported in the Table 8. The features that are discarded are discussed in Section 6. The incremental improvement in the system performance in both languages for each of the features is listed in Table 9. The graphical representation of the incremental improvement in the system performance as features are added is shown in Figure 2 for both languages. It may be observed that positional feature and the average distribution feature plays very crucial role to identify sentence subjectivity. The evaluation results of the final system after inclusion of all the features listed in Table 8 have been shown in Table 10 for both the languages. The evaluation results of the rule based theme detection technique are clearly an improvement over the evaluation results of the baseline systems on the MPQA corpus. Later on MPQA subjectivity word list [36] has been added as ontology knowledge for English and the overall system performance increased by 1.89%. We are now working to develop a subjectivity word list for Bengali for future task. The evaluation result of Subjectivity detection on English with MPQA subjectivity word list is shown in Table 11.

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

Table 8. Feature Set

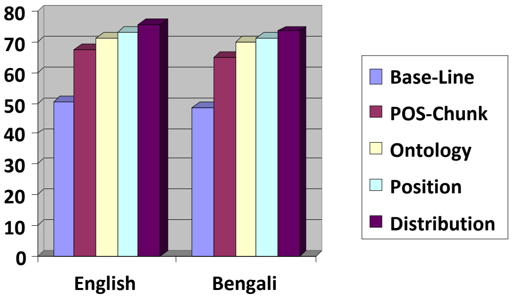

Figure 2. Feature wise System Performance

5. Conclusion and Future Work

In this paper we describe a prototype rule based system for identifying subjectivity through Theme Detection for English and Bengali. A number of features are identified for theme detection and detailed experiments are carried out to identify the role of each feature. Each sentence in the corpus is identified as opinionated or non-opinionated sentence. Evaluation of the system has yielded satisfactory results.

We are now working on improving the performance of the present system. Future task will be in the direction of classifying Theme knowledge in polarity classes such as positive, negative and neutral, development of techniques for creation of opinion summaries and tracking of opinions over a period of time on a particular subject-aspect combination among others.

References

[1] Wiebe, Janyce M.and Wilson Theresa and Bruce, Rebecca F. and Bell, Matthew and Martin, Melanie. Learning Subjective Language. In proceeding of Computational linguistics. 277—308.

[2] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[3] Ellen Riloff, Janyce Wiebe and William Phillips. Exploiting subjectivity classification to improve information extraction. In Proceedings of AAAI, pages 1106–1111, 2005.

[4] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[5] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (ACL), pages 417–424, 2002.

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

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

[8] 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.

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

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

[11] 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.

[12] 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.

[13] Hu and Liu. Mining and summarizing product re-views. Preoceedings of 10th ACM SigKDD, 2004.

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

[15] Esuli A and Sebastini F. Determining the semantic orientation of terms through gloss analysis. Procesd-ings of CIKM, 2005.

[16] Yejin Choi, Clarie Cardie, Ellen Riloff and Sidd-harth 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.

[17] Andrew Smith, Trevor Cohn and Miles Osborne. Logarithimic Opinion Pools for Conditional Random Fields. In Proceeding of the 43rd Annual Meeting of the ACL, pages 18-25, 2005.

[18] 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.

[19] Gregory Grefenstette, Yan Qu, James G. Shanahan and David A. Evans. Recherche d'Information As-sistée par Ordinateur. In Proceedings of RIAO, 7th International Conference on 2004.

[20] 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.

[21] 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.

[22] 1A. 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.

[23] A. Esuli and F. Sebastiani. PageRanking WordNet synsets: An application to opinion mining. In Proceedings of the Association for Computational Linguistics (ACL), 2007.

[24] Esuli and Sebastini. Sentiwordnet: a publicly available resource for opinion Genova, Italy. 2006.

[25] Nathan Eagle, Push Singh and Alex (Sandy) Pent-land .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).

[26] Ekbal, A., Bandyopadhyay. S. A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation Journal. pages 173-182, 2008

[27] Janyce Wiebe, Theresa Wilson and Claire Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation (formerly Computers and the Humanities) 1(2), 2005.

[28] Janyce M. Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from un-annotated texts. In Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing), number 3406 in Lecture Notes in Computer Science, pages 486–497, 2005.

[29] 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.

[30] 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.

[31] 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.

[32] Dan Klein and Christopher D. Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pages 3-10, 2003.

[33] J. Lafferty, A. K. McCallum and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of 18th International Conference on Machine Learning. 2001.

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

[35] Karen Kukich . Techniques for automatically correcting words in text. ACM Computing Surveys, 24(4), pages 377-439, (1992).

[36] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.