Sentiment Analysis of Hindi Review based on Negation and Discourse Relation

Namita Mittal
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
mittalnamita@gmail.com

Basant Agarwal
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
thebasant@gmail.com

Garvit Chouhan
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
jkgarvit@gmail.com

Nitin Bania
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
nittinuts@gmail.com

Prateek Pareek
Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur
prtkpareek@gmail.com

Abstract: With recent developments in web technologies, percentage of web content in Hindi language is growing up at a lightning speed. Opinion classification research has gained tremendous momentum in recent times mostly for English language. However, there has been little work in this area for Indian languages. There is a need to analyse the Hindi language content and get insight of opinions expressed by people and various communities. In this paper, a method is proposed to increase the coverage of the Hindi SentiWordNet for better classification results. In addition to this, impact of the negation and discourse rules are investigated for Hindi sentiment analysis. Proposed algorithm produces 82.89% for positive reviews and 76.59 % for negative reviews, and an overall accuracy of 80.21%.

Keywords: Sentiment Analysis, HSWN, Discourse and negation for Hindi Reviews.

1. Introduction

Sentiment Analysis is a natural language processing task that deals with the extraction of opinion from a piece of text with respect to a topic (Pang et al., 2008). A large number of advertising industries and recommendation systems work on understanding liking and disliking of the people from their reviews. Hindi is the fourth highest speaking language in the world. The increasing user-generated content on the Internet is the motivation behind the sentiment analysis research. Majority of the existing work in this field is for English language. Very little attention has been paid in direction of sentiment analysis for Hindi Language. Information content in Hindi is important to be analysed for the use of industries and government(s).

Sentiment analysis is very difficult for Hindi language due to numerous reasons as follows. (1) Unavailability of well annotated standard corpora, therefore supervised machine learning algorithms cannot be applied. (2) Hindi is a resource scarce language; there are no efficient parser and tagger for this language. (3) Limited resources available for this language like HindiSentiWordNet (HSWN). It consists of limited numbers of adjectives and adverbs. All the words are available in inflected forms. Even all the inflected forms of the word are not present. HSWN is created using the Hindi WordNet and English SentiWordNet (SWN). During the creation of this resource for Hindi language, it is assumed that all synonyms have the same polarity while all antonyms have the reverse polarity of a word. This assumption neglected word sense intensity in terms of polarity, however polarity intensity of their word is important in opinion mining. (4) Even, Translation dictionaries may not account for all the words because of the

International Joint Conference on Natural Language Processing, pages 45–50,
Nagoya, Japan, 14-18 October 2013.

language variations. Same words may be used in multiple contexts and context dependent word mapping is a difficult task, error prone and requires manual efforts. Using Translation method for generating subjective lexicon, there is a high possibility of losing the contextual information and sometimes may have translation errors.

In this paper, an efficient approach is proposed for identifying sentiments and opinions from user generated content in Hindi.

Main contributions of this paper are as follows. (1) Developed an annotated corpus for Hindi Movie Reviews. (2) Improve the existing HindiSentiWordNet (HSWN) by incorporating more opinion words into it. (3) Proposed new rules for negation handling and discourse relation for Hindi language reviews.

This paper is organised as follows. Section 2 presents related work. Proposed approach is described in detail in Section 3. Section 4 discusses the experimental setup and results. Finally, Section 5 concludes and presents the future work.

2. Related Work

Identifying the sentiment polarity is a complex task. To address the problem of sentiment classification various methods have been proposed (Agarwal et al. 2012, Agarwal et al. 2013, Pang et al. 2008). Joshi et al. (2010) proposed a fallback strategy in their paper. This strategy follows three approaches: In-language Sentiment Analysis, Machine Translation and Resource Based Sentiment Analysis. The final accuracy achieved by them is 78.14 %. They developed a lexical resource, HindiSentiWordNet (HSWN) based on its English format. Bakliwal et al. (2012) created lexicon using a graph based method. They explored how the synonym and antonym relations can be exploited using simple graph traversal to generate the subjectivity lexicon. Their proposed algorithm achieved approximately 79% accuracy on classification of reviews and 70.4% agreement with human annotated. Mukherjee et al. (2012) showed that the incorporation of discourse markers in a bag-of-words model improves the sentiment classification accuracy by 2 - 4%. Bakliwal et al. (2011) proposed a method to classify Hindi reviews as positive or negative. They devised a new scoring function and test on two different approaches. They also used a combination of simple N-gram and POS-Tagged N-gram approaches. Ambati et al. (2011) proposed a novel approach to detect errors in the treebanks. This approach can significantly reduce the validation time. They tested it on Hindi dependency treebank data and were able to detect 76.63% of errors at dependency level.

3. Proposed Approach

Proposed approach for Sentiment Analysis of Hindi review documents works as follows. Initially, annotated dataset is created for testing of the proposed algorithm. Then, rules are devised for handling negation and discourse relation which highly influence the sentiments expressed in the review. Further, HindiSentiWordNet (HSWN) is used for polarity values of words. Method for improving the HSWN is also proposed. Finally, overall semantic orientation of the review document is determined by aggregating the polarity values of all the words in the document

3.1. Preparation of Annotated Dataset

Initially, 900 reviews are crawled from Hindi review websites, out of these 900 reviews, 130 reviews were rejected due to their objective nature manually. Next, for remaining 770 reviews, agreement was established on 662 reviews using Cohen's kappa. Out of these 662 total reviews, 380 were agreed as positive and 282 as negative. After that, Fleiss kappa was used for the agreement and achieved 0.8092 as kappa coefficient. This falls under the substantial agreement according to Fleiss kappa. Average size of the reviews in our dataset is 104 words.

3.2 Negation Handling

The negation operator (Example: नही, न, नदारद etc.) inverts the sentiment of the word following it. The usual way of handling negation in sentiment analysis is to consider a window of size n (typically 3 to 5) and reverse the polarity of all the words in the window.

We reverse all the words in the window by adding (!) to every word, till either the sentence is completed or a violating expectation (or a contrast) conjunction or a delimiter is encountered. Negation on the basis of sentence structure may be applied either in forward or in backward direction. Some rules are proposed to handle negation, are discussed in following cases.

CASE 1: If a sentence has only one single negate word ("नही", "नदारद") i.e. negation is present in a simple sentence. For example.

(1) यह मूवी अच्छी नही हैं । (2) कागज पर लिखी गई कहानी का ठीक से फिल्मी रूपांतरण नहीं किया गया है |

In the above sentence, due to negation, all the words before the negation word "नही" would be negated and the reverse polarity of the negated words would be considered further. The above examples will be negated as

(1) !यह !मूवी !अच्छी नही हैं ।

(2) !कागज !पर !लिखी !गई !कहानी !का !ठीक !से !फिल्मी !रूपांतरण नहीं किया गया है

But this negation rule may be invalid for sarcastic and special form of sentences.

e.g. कोई भी मूवी इससे बढ़िया नही हो सकती ।

CASE 2: If a sentence has a negate word and conjunction, and index of conjunction is more than the index of negated word, forward negation is applied. For example:

(1) पूरी फिल्म इस तरह की नहीं बन पाई कि आम आदमी उसे पूरे समय रुचि से देखे|

(2) कॉमेडी फिल्म होने के बावजूद इसमें ऐसा कुछ भी नही जो दर्शकों को हँसा सके ।

In the above sentences, negate word and the conjunction words are present and the index of conjunction is greater than the index of negate word; therefore, forward negation is applied. In above example, all the words after the conjunction will be negated .The above examples will be negated as

a) पूरी फिल्म इस तरह की नहीं बन पाई कि !आम !आदमी !उसे !पूरे !समय !रुचि !से !देखे|

b) कॉमेडी फिल्म होने के बावजूद इसमें ऐसा कुछ भी नही !जो !दर्शकों !को !हँसा !सके ।

CASE 3: If a sentence has "न" multiple times in sub-sentences separated by commas. For example: (1) न कहानी ढंग की है, न पटकथा और न ही निर्देशन|

"न" usually occurs multiple times in this example sentence, with sub sentences separated by commas. Here for each "न" the negation is applied in forward direction until a delimiter is encountered. The above example will be negated as follows. न !कहानी !ढंग !की !है, न !पटकथा और न !ही !निर्देशन|

3.3 Discourse Relations

An essential phenomenon in natural language processing is the use of discourse relations to establish a coherent relation, linking phrases and clauses in a text. The presence of linguistic constructs like connectives, modals, and conditional can alter sentiment at the sentence level as well as the clausal or phrasal level (Wolf et al., 2005). A coherent relation reflects how different discourse segments interact. Discourse segments are non-overlapping spans of text. In this paper, Violated Expectations like हालाकि, लेकिन, जबकि etc. are handled.

Violating expectation conjunctions oppose or refute the neighboring discourse segment. These conjunctions are categorized into the following two sub-categories: Conj_After and Conj_Infer.

3.3.1 Conj_After:

It is the set of conjunctions that give more importance to the discourse segment that follows them. It means that actual segment is mostly reflected by the statement following the conjunction. So, in all the below examples, the discourse segments after the Conj_After (in bold) are given preferences and the previous sentences are dropped.

For example: लेकिन , मगर , फिर भी, बावजूद

लेकिन: कहने को तो फिल्म दो घंटे की हैं, लेकिन ये दो घंटे किसी सजा से कम नहीं है|

मगर: फिल्म कई जगह चमक छोड़ती है मगर कुल मिलाकर बात बन नहीं पाती |

बावजूद: इतने सारे संसाधन होने के बावजूद साबिर मनोरंजक फिल्म नहीं बना सके|

फिर भी: वैसे तो इस फिल्म में ऐसा कुछ नहीं है जो दर्शकों को आकर्षित कर पाए ,फिर भी विवेक ऑबराय की कॉमेडी देखने के लिए दर्शक थियेटर की ओर रुख कर सकते हैं|

3.3.2 Conclusive or Inferential Conjunctions

These are the set of conjunctions, Conj_infer, that tend to draw a conclusion or inference. Hence, the discourse segment following them should be given more weight.

For example: इसीलिए , कुल मिलाकर

कुल मिलाकर : कुल मिलाकर 'ब्रेक के बाद' ब्रेक से पहले ही अच्छी है |

3.4 Improvement of HSWN

Existing version of HindiSentiWordNet consists of limited numbers of adjectives and adverbs. All those words are available in inflected forms. Even all the inflected forms of the word are not present. HSWN is created using the Hindi WordNet and English SentiWordNet (SWN). During the creation of this resource for Hindi language, it is assumed that all synonyms have the same polarity while all antonyms have the reverse polarity of a word. HSWN is improved in the same way as it was developed initially. The main focus during the improvement was on missing and inflected adjectives and adverbs. Therefore, all the inflected words of the existing root words are also included in the improved HSWN. Proposed approach is describes in Algorithm 1. In Step 4, Google translator is used in our experiment. In Step 6, in case of sense disambiguation, the suitable sense of the word refers to the sense which is suitable according to the domain.

Algorithm 1. Improvement of HSWN

Step 1: Find out the adjectives and adverbs in the corpus that are not in HSWN.
Step 2: Extract adjectives and adverb from document corpus.
Step3: Now for each of the extracted word in Step 2.
Step 4: Translate the given word into its English meaning using a bilingual resource.
Step 5: Find the polarity of the translated word using English SentiWordNet. If single entry is found then go to step7.
Step 6: Select the entry with the suitable and most common sense of the word.
Step 7: Translate the word back to Hindi
Step 8: Add it to the HSWN
Step 9: return

In our case the domain is the movie review dataset. If multiple senses are possible in the same domain, then select the most common sense among these words, which implies that multiple resources may need to be created for different domains.

3.5 Proposed Algorithm for Sentiment Analysis of Hindi Reviews

The first step of the proposed algorithm is the pre-processing.

Algorithm 2. Proposed Algorithm

Step 1: For each document in the corpus
Step 2: Apply Pre-Processing
(a) Remove the Stop Words.
(b) Apply Rules (Negation and Discourse).
End of For Loop of Step 1;
Step 3: For each token in the document.
Step 4: Retrieve polarity (POL) from modified HSWN.
Step 5: **If** (word is present in HSWN)
Then go to Step 6
Else Add it to Missing Word List
Step 6: **If** (word is negated)
Then word.POL=-POL;
Else Word.POL=POL;
End of For Loop of Step 3;
Step 7: Compute the aggregate polarity of the document (doc.POL) by adding the polarities values of all the token.
Step 8: **If** (doc.POL > zero)
Then label the document as positive
Else If (doc.POL<zero)
Then label the document as negative
Else Classify the document as neutral.
Step 9: Return the set of Labelled Documents.

Review documents are pre-processed by applying stemming, negation and discourse relations as discussed in previous sections. After, the pre-processing, polarity value is retrieved from the improved HindiSentiWordNet (HSWN). Finally, by aggregating the polarity values of all the words semantic orientation of the review document is determined. Proposed approach is describes in Algorithm 2.

4. Results and Discussions

Proposed algorithm is tested on 662 movie review dataset created as described in previous sections. For various experimental settings, results are reported in Table 1. Initially, semantic orientation of a document is determined by aggregating the total polarity value of all the words in the document using existing HSWN. Experimental results show an accuracy of 50.45%, which is very less. The main reason for this observation was that most of the words in our dataset were not present in the HSWN and some words are inflected forms of the available words in HSWN. Further, proposed algorithm without any negation and discourse handling is applied using improved HSWN, and experimental results show that accuracy increased up to 69.79%. The proposed algorithm performs well for positive reviews, for the negatives performance needs to be improved.

Table 1. Accuracy of various experiments

		ACCURACY (In %)		
S. No.	Experimental Setup	Positive	Negative	Overall
1	Only Existing HSWN	50	51.06	50.45
2	With Improved HSWN	85.26	48.93	69.78
3	With Improved HSWN + Negation	82.89	72.34	78.39
4	Improved HSWN +Negation+ Discourse	82.89	76.59	80.21

In our further versions of the experiments, we analysed the impact of our negation rules and applied proposed algorithm with negation on the movie review dataset. Experimental results show an improvement in performance for overall sentiment analysis especially for the negative reviews. Overall accuracy with negation handling increased to 78.39 %. Further, to we applied discourse relation with negation rules on reviews, and experimental results show that significant improvement for sentiment classification. Results obtained for positive, negative and total reviews are 82.89%, 76.59% and 80.21% respectively.

5. Conclusion and Future Work

Opinion Mining for Hindi is an important task. In the paper, a method is proposed to increase the coverage of HindiSentiWordNet (HSWN) for better classification results, as HSWN faces the problem of very less coverage. In addition to this, impact of negation and discourse are investigated on Hindi Review sentiment analysis. This approach just uses only one resource HSWN for the word polarity. The movie review corpus is developed in Hindi using the Hindi websites as our source. It has been standardized using Cohen's Kappa and Fleiss Kappa for agreement. Improvement of HSWN is proposed for improved results. The inflected forms of the existing root words in this HSWN are also included. Experimental results show that proposed algorithm with negation and discourse relations achieves 82.89% for positive reviews and 76.59 % for negative reviews with an overall accuracy of 80.21%. In future, the dataset can further be extended for the better and generalized results. This work can be extended to incorporate Word Sense Disambiguation (WSD) and morphological variants which could result in better accuracy for words which have dual nature. HSWN may be developed further.

References

Aditya Joshi, Balamurali A R, Pushpak Bhattacharyya. "A Fall-Back Strategy For Sentiment Analysis In Hindi: A Case Study", In International Conference On Natural Language Processing (ICON), 2010.

Akshat Bakliwal, Piyush Arora, Vasudeva Varma. "Hindi Subjective Lexicon : A Lexical Resource For Hindi Polarity Classification".In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC) 2012.

Akshat Bakliwal, Piyush Arora, Ankit Patil, Vasudeva

Varma, "Towards Enhanced Opinion Classification using NLP Techniques" In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 101–107, 2011

Basant Agarwal, Namita Mittal, "Optimal Feature Selection Methods for Sentiment Analysis", In 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013),Vol-7817, pages-13-24, 2013

Basant Agarwal, Namita Mittal, "Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification", In Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), COLING 2012, pages 17–26, 2012.

Bharat R. Ambati, Samar Husain, Sambhav Jain, Dipti M. Sharma, Rajeev Sangal, "Two Methods to Incorporate Local Morph Syntactic Features in Hindi Dependency Parsing" In Proceedings of the NAACL HLT 1st Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 22–30, 2010.

Florian Wolf and Edward Gibson. "Representing Discourse Coherence: A Corpus-based Study". Computational Linguistics, 31(2), pp. 249-287. 2005

Bo Pang, Lillian Lee, "Opinion mining and sentiment analysis". Foundations and Trends in Information Retrieval, Vol. 2(1-2):pp. 1–135. (2008).

Subhabrata Mukherjee, Pushpak Bhattacharyya, "Sentiment Analysis in Twitter with Lightweight Discourse Analysis", In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), 2012