Discourse Based Sentiment Analysis for Hindi Reviews

Namita Mittal, Basant Agarwal, Garvit Chouhan, Prateek Pareek, and Nitin Bania

Department of Computer Engineering,
Malaviya National Institute of Technology, Jaipur, India
nmittal@mnit.ac.in,
{thebasant, jkgarvit, prtkpareek, nitinnuts}@gmail.com

Abstract. Research on Sentiment Analysis (SA) has increased tremendously in recent times due to fast growth in Web Technologies. Hindi Language content is also growing very fast online. Sentiment classification research has been done mostly for English language. However, there has been little work in this area for Indian languages. Sentiment analysis means to extract the opinion expressed in the text about a specific topic. There is a need to analyse the Hindi language content and get insight of opinions expressed by people and various communities about a specific topic. In this paper, it is investigated that how by proper handling of negation and discourse relation may improve the performance of Hindi review sentiment analysis. Experimental results show the effectiveness of the proposed approach.

Keywords: Sentiment Analysis, HSWN, Discourse relations, negation handling, Hindi Reviews.

1 Introduction

Sentiment Analysis is a natural language processing task that deals with the findings of opinion in a piece of text with respect to a topic [9]. A large number of advertising industries and recommendation systems work on understanding the people likings and disliking's from this content. 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.

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 not 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. Even, most of the words are available in inflected forms. Also, 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 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 objective of this paper is to investigate the influence of negation handling and discourse relations on the performance of Hindi review sentiment analysis.

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

To identify the sentiment expressed in the text is difficult task for Hindi language. A lot of work has been done on sentiment analysis has been done mostly for English language [3], [5], [9], but for Hindi, sentiment analysis research in initial phase. In [2], authors 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. In [1], authors 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, HSWN based on its English counter format. In [6], authors investigated the use of discourse and negation with the enhancement of the HSWN for Hindi reviews. In [7], authors showed that the incorporation of discourse markers in a bag-of-words model for English language improves the sentiment classification accuracy by 2 - 4%. In [4], authors 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.

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. Some basic rules are devised for negation and discourse handling which highly influence the sentiments expressed in the review. Further, HindiSentiWordNet

(HSWN) is used for the polarity values of words. Finally, overall semantic orientation of the review document is determined by aggregating the polarity values of all the words present 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.

e.g. (1) **इस मूवी का निर्देशन अच्छा नहीं हैं ।** (2) **मूवी की कहानी में दम नहीं हैं ।**

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) **!इस !मूवी !का !निर्देशन !अच्छा नहीं हैं ।** (2) **!मूवी !की !कहानी !में !दम नहीं हैं ।**

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

e.g. **इससे बढ़िया एक्टिंग हो ही नहीं सकती ।**

CASE 2: If a sentence has a negation word and conjunction, and index of conjunction is more than the index of negated word, forward negation is applied. For example: (1) **फिल्म की कहानी ऐसी नहीं हैं कि इसे तीन घंटों तक मजे से देखा जा सके ।** (2)

बढ़िया एक्टिंग के बावजूद भी कहानी में ऐसा कुछ भी नहीं जो दर्शकों को बांधे रख सके । In these 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 follows. (1) **फिल्म की कहानी ऐसी नहीं हैं कि !इसे**

!तीन !घंटो !तक !मजे !से !देखा !जा !सके । (2) बढ़िया एक्टिंग के बावजूद भी कहानी में ऐसा कुछ भी नहीं जो !दर्शको !को !बांधे !रख !सके ।

CASE 3: If a sentence have “न” multiple times in sub-sentences separated by commas. For example: (1) **न एक्टिंग सही हैं , न मूवी की कहानी ।**

“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 !एक्टिंग !सही ! हैं , न !मूवी !की !कहानी । ।**

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 [8]. 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: फिल्म की कहानी ठीक हैं, लेकिन खराब एक्टिंग से बात बिगड़ गई ।

मगर: फिल्म इंटरवल के बाद ठीक हैं ,मगर कुल मिलाकर वो बात नहीं बन पाई ।

बावजूद: अच्छे डायरेक्शन के बावजूद भी फिल्म अपना प्रभाव नहीं बना पाई ।

फिर भी: वैसे मूवी औसत हैं , फिर भी एक बार देखी जा सकती 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 यह मूवी समय की बर्बादी हैं ।

3.4 Proposed Algorithm for Sentiment Analysis of Hindi Reviews

The first step of the proposed algorithm is the pre-processing. Review documents are pre-processed by applying stemming, negation and discourse relations as discussed in previous sub-sections. After, the pre-processing, polarity values are retrieved from the HSWN. Finally, semantic orientation of the overall review document is determined by aggregating the polarity values of all the words. Proposed approach is describes in Algorithm 1.

Algorithm 1. Proposed Algorithm

```

Step 1: For each document in the corpus
Step 2: Apply Pre-Processing
        (a) Remove the Stop Words and apply Stemming.
        (b) Apply Rules (Negation and Discourse).
Step 3: For each token in the document.
Step 4: Retrieve polarity (POL) from HSWN.
Step 5: If (word is negated)
        Then word.POL=-POL;           Else   Word.POL=POL;
Step 6: Compute the aggregate polarity of the document (doc.POL) by adding the
        polarities values of all the token.
Step 7: 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 8: Return the set of Labelled Documents
  
```

4 Results and Discussions

Proposed algorithm is tested on 662 movie review dataset created by our own as described in previous section. For various experimental settings, results are reported in Table 1.

Table 1. Accuracy of various experiments

S. No	Experimental Setup	Accuracies (In %)		
		Positive	Negative	Overall
1	With only HSWN	50	51.06	50.45
2	HSWN + Negation	71.32	79.71	74.92 (+48.5%)
3	HSWN + Discourse	78.90	71.33	75.67 (+49.9%)
4	HSWN + Negation +Discourse	81.86	75.54	79.15 (+56.8%)

First of all, Semantic orientation of a document is determined by aggregating the total polarity value of all the words in the document using HSWN. Experimental results show an accuracy of 50.45%, which is very less. This accuracy is considered as baseline accuracy. 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 is experimented with negation rules; it produces accuracy of 74.92% (+48.5%). Negation rules applied produces significant improvement over baseline accuracy. The main improvement due to negation was in negative reviewed documents. Further, impact of discourse relation is experimented, which gives an accuracy of 75.67% (+49.9%). Further, both negation and discourse rules are applied; it gives an accuracy of 79.15% (+56.8%).

5 Conclusion and Future Work

Opinion Mining for Hindi is an important task. In this paper, it is investigated that how the negation and discourse relations can be efficiently handled for improving the performance of sentiment analysis for Hindi reviews. Proposed approach uses the resource HSWN for the word polarity. The movie review corpus is developed in Hindi Language from the Hindi review websites. Experimental results show that proposed algorithm with negation and discourse relations significantly improves the performance for sentiment analysis. 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

1. Joshi, A.R., Balamurali, P.: A Fall-Back Strategy For Sentiment Analysis In Hindi: A Case Study. In: International Conference on Natural Language Processing, ICON (2010)
2. Bakliwal, A., Arora, P., Varma, V.: Hindi Subjective Lexicon: A Lexical Resource For Hindi Polarity Classification (2012)
3. Agarwal, B., Mittal, N.: Optimal Feature Selection Methods for Sentiment Analysis. In: Gelbukh, A. (ed.) CICLing 2013, Part II. LNCS, vol. 7817, pp. 13–24. Springer, Heidelberg (2013)
4. Bakliwal, P., Arora, A., Patil, V.: Towards Enhanced Opinion Classification using NLP Techniques. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pp. 101–107 (2011)
5. Agarwal, B., Mittal, N.: 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, COLING 2012, pp. 17–26 (2012)
6. Mittal, N., Agarwal, B., Chouhan, G., Pareek, P., Bania, N.: Sentiment Analysis of Hindi Review based on Negation and Discourse Relation. In: 11th Workshop on Asian Language Resources (ALR), In Conjunction with IJCNLP (in press)
7. Mukherjee, S., Bhattacharyya, P.: Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In: Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012 (2012)
8. Wolf, F., Gibson, E.: Representing Discourse Coherence: A Corpus-based Study. Computational Linguistics 31(2), 249–287 (2005)
9. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)