A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon

Raksha Sharma, Pushpak Bhattacharyya
Dept. of Computer Science and Engineering
IIT Bombay, Mumbai, India
{raksha,pb}@cse.iitb.ac.in

Abstract

Supervised approaches have proved their significance in sentiment analysis task, but they are limited to the languages, which have sufficient amount of annotated corpus. Hindi is a language, which is spoken by 4.70% of the world population, but it lacks a sufficient amount of annotated corpus for natural language processing tasks such as Sentiment Analysis (SA). With the increase in demand and availability of Hindi review websites, an accurate sentiment analyzer for Hindi has become a need.

In this paper, we present a bootstrap approach to extract senti words from HindiWordNet. The approach is designed such that it minimizes the extraction of the words with the wrong polarity orientation, which is a crucial task, because a word can have positive and negative senses at the same time. The resultant set of 8061 polar words, we call it Hindi senti lexicon, is used for sentiment analysis in Hindi. We get an average accuracy of 87% for sentiment analysis in the movie and product domain.

1 Introduction

In the real world, people find themselves comfortable in their national language, both in case of reading and writing. Hindi is the national language of India, spoken and understood almost all over the country. Keeping this in mind, there is a tremendous growth in the Hindi review websites[1] on the Web. Besides this, a few Hindi lovers like to post their reviews in Hindi on English based e-commerce websites also, for example, we can find Hindi reviews on *www.flipkart.com* or *www.homeshop18.com* in a big number with English reviews. Therefore, an efficient sentiment analysis system for the Hindi language is the need of current e-commerce organizations and their customers.

The general approach of Sentiment Analysis (SA) is to summarize the semantic polarity (i.e., positive or negative) of sentences/documents by analysis of the orientation of the individual words (Riloff and Wiebe, 2003; Pang and Lee, 2004; Danescu-Niculescu-Mizil et al., 2009; Kim and Hovy, 2004; Takamura et al., 2005). In the absence of sufficient amount of corpora, the most efficient way of sentiment analysis is to rely on the words from sentiment lexicons as a key feature. There are many sentiment lexicons for English language, for example, subjectivity lexicon[2] by Wiebe and a list of positive and negative opinion words[3] by Liu, but there are not many instances of sentiment lexicons in Hindi that can build an efficient sentiment analysis system for Hindi. The main contributions of this paper are:

- A Hindi senti lexicon consisting polar words of four parts of speech: Adjective, Noun, Verb and Adverb, generated from extensive analysis of HindiWordNet[4].
- A multi module rule based sentiment analysis system for Hindi that uses words from Hindi senti lexicon as a polarity clue.

Our approach that generates Hindi senti lexicon is an improvement over the approach suggested by Bakliwal et al. (2012). They used the same source, that is, HindiWordNet for the polar words extraction, but their approach was not able to handle the instances, where a word has senses of both the orientations: positive and negative. A word can have

[1] hindi.webdunia.com/entertainment/film/review/

[2] http://mpqa.cs.pitt.edu/

[3] http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html

[4] Available at: www.cfilt.com

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 150–155,
Goa, India. December 2014.

positive and negative senses at the same time. This combination is exemplified in figure 1. The approach of Bakliwal et al.(2012) assumes that all the senses (synsets) of a source word should receive the polarity same as the source word, consequently all the words in the synsets should bear the polarity same as the source word. This assumption will fetch the words with the wrong polarity orientation. Besides this, their approach considers that antonyms of a word should receive the opposite polarity of the source word, but it is not always true. There are the instances, where antonym and the word have the same polarity orientation. One such instance of this property is given in figure 2. In this paper, we present an approach that provides a very vigilant traversal of the HindiWordNet for the purpose of generation of senti lexicon. Our approach also implements that synonyms (words from a synset) must receive the same polarity, but it is able to distinguish among the positive and negative senses of the same word.

Index word = सस्ता with +ve polarity orientation	Synsets of सस्ता with actual Polarity orientations
Sense-1	+ve (सस्ता, अल्पमूल्य, सौघा, अनर्घ, सुहंग, अल्पकरीत; जो कम मूल्य का हो)
Sense-2	-ve (घटिया, निकृष्ट, नीच, ओछा, छिछोरा, तुच्छ, कमीना etc; बिल्कुल निम्न या निकृष्ट कोटि का)

Figure 1: Synsets of the same word with opposite polarities

Results show that the sentiment analysis system based on sentiment lexicon generated by the proposed bootstrap approach outperforms the Hindi sentiment analysis system presented by Bakliwal et al.(2012) for the same data set. By considering four parts of speech as a source of polar words, our sentiment lexicon covers more range of polar words. We achieve an accuracy of 89.45% in the product domain and 85% in the movie domain.

Index word = सस्ता with +ve polarity orientation	Antonym of सस्ता with positive polarity orientation
+ve (सस्ता, अल्पमूल्य, सौघा, अनर्घ, सुहंग, अल्पकरीत - जो कम मूल्य का हो)	+ve (बहुमूल्य, बेशकीमती, बेशक़ीमती, मूल्यवान, क़ीमती, कीमती, अनमोल - जिसका मूल्य बहुत अधिक हो)

Figure 2: Antonyms with the same polarity as the source/index word

Section 2 helps illustrate the generation of senti lexicon from HindiWordNet. Section 3 provides the statistics related to sentiment lexicon obtained as output. Section 4 expands on the the rule based classification algorithm that is used to find the overall polarity orientation of the document. Section 5 illustrates the results achieved for Sentiment analysis system. Sections 6 and 7 discuss related work and conclusion.

2 Identification of Polar Words from HindiWordNet

All the senses of a word may not necessarily be polar. In this section, we focus on the extraction of polar words from HindiWordNet.

2.1 Why HindiWordNet?

We find HindiWordNet the most efficient resource for generation of senti lexicon. HindiWordNet is the result of manual identification of all the senses of the word W, hence it is very accurate in terms of relations among the words. We give a formal characterization of HindiWordNet. A WordNet is a word sense network. A word-synset network N is a triple (W, S, ε), where W is a finite set of words, S is a finite set of synsets, ε is a set of undirected edges to link synsets of words, that is, $\varepsilon \subseteq W \times S$. Each synset in the Hindi WordNet is linked with other synsets through the well-known semantic relations of hypernymy, hyponymy, meronymy, troponymy, antonymy, entailment *etc*.

In HindiWordNet, the words are grouped in a synset according to a lexical concept, that is, synonymy. Two words that can be interchanged in a context are synonymous in that context. We observe that the synonymy relation is the most specific candidate for lexicon generation. The other semantic relations fetch the non polar words on expansion. This synonymy property assures that if we know the polarity orientation of a word in the synset, the same polarity orientation can be assigned to all the words in the synset. Figure 3 exemplified the inference of polarity orientation from a seed (index) word to the whole synset.

2.2 A Bootstrap Approach for Polar Words Identification

The process is based on the phenomenon that if we know the polarity orientation of a word, the same polarity orientation can be assigned to a synonymous word. Therefore, we start with two seed sets: positive (Each word has polarity value: +1) and negative (Each word has polarity value: -1). The

Index word with polarity orientation	Synset with inferred polarity orientation from the index word
-ve (घटिया)	-ve (घटिया, निकृष्ट, नीच, ओछा, छिछोरा, तुच्छ, कमीना, बाज़ारू, बाज़ारी, बज़ारू, बाजारू, बाजारी *etc.*)
+ve (विनम्र)	+ve (विनम्र, विनयी, विनीत, नम्र, विनयशील, विनययुक्त, आनत, निभृत, अनुनीत, प्रवण, अवाग्र, आजिज़, आजिज)

Figure 3: Polarity identification from a seed word to the whole synset

seed (index) words are manually identified polar words.

For each index word from seed set, we extract all the senses, considering they all will have the same polarity orientation as the index word, but there are instances where a word can have both positive and negative senses. To assure the rejection of opposite polarity sense, we extract all the words that belong to the sense and check the presence of any word in the opposite polarity seed-set or lexicon. For example, for a positive index word, we look into the negative seed-set or lexicon. For the first iteration, only seed-set is available for this check, in further iterations it becomes lexicon. If any word from the extracted sense is found in the opposite polarity lexicon, discard the sense and repeat the process for the next sense of the index word, else insert the words into the lexicon having polarity orientation of index word.

The decision about the discarded sense will remain pending till the word from the sense, which is found in the opposite polarity lexicon encounters as an index word. The whole process will be repeated for the newly added words till no new word is added in the lexicons. Once the process stops, it results into two separate lexicons: positive and negative lexicons. We combine these two as *Hindi senti lexicon*. The whole process is depicted in figure 4.

Objective Sense (synset): A polar index word may also have objective sense. To minimize the extraction of objective synsets, we analyzed the behavior of the words in a synset comprehensively. The words in a synset in HindiWordNet are arranged in the order of their frequency of usage. The words, which are at the head of the synset (most frequently used) fetch the polar synsets on further expansions, but the words which are at the tail of the synset (less frequently used) are prone to fetch objective synsets on further expansion. The

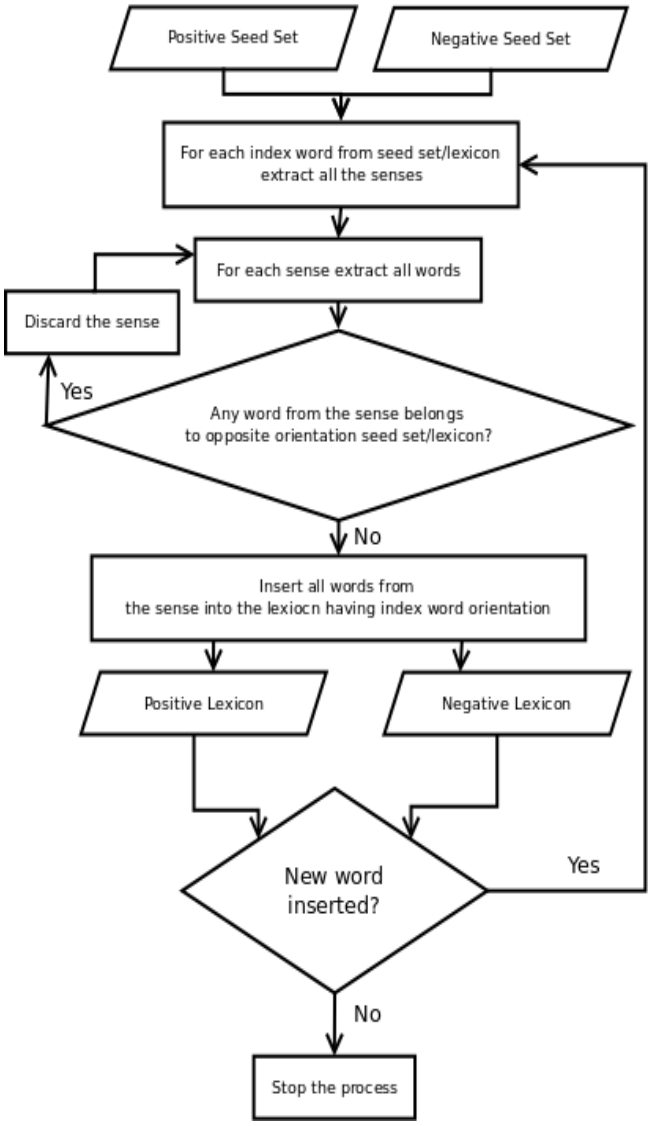

Figure 4: A bootstrap approach for polar words identification

head words are tightly coupled with polar synsets, while tail words fetch the polar synsets that can be extracted from head words, besides this they extract non polar synsets. To hinder the extraction of non polar words, we add a restriction that only the first seven (head) words of a synset will take part as index word in the next iteration, and the rest will remain the part of lexicon only. This constant 7 is selected after experimenting with a range of constants 2 to 10, such that it minimizes the extraction of the non polar words as well as keeps the lexicon compact.

3 Hindi Senti Lexicon Statistics

Most of the sentiment lexicons are limited to adjectives and adverbs, but there are many instances where a noun and verb can bear the polarity. We apply the same bootstrap approach discussed in section 2.2 on all four parts of speech, which are prone to bear polarity. The proposed constrained approach minimizes the possibility of extraction of wrong word as polar word. We extracted a comprehensive list of total 8061 true polar words, which we name as Hindi senti lexicon. The efficacy of this lexicon is evaluated by implementing Hindi sentiment analysis system (section 4) in the movie and product domains.

	Number
Adjective	Pos:2256 & Neg:2232
Noun	Pos:1551 & Neg:1558
Adverb	Pos:132 & Neg:94
Verb	Pos:83 & Neg:155
Total	8061

Table 1: Number of polar words obtained in different parts of speech

4 A Multi-module Sentiment Analysis System for Hindi Using Hindi Senti Lexicon

The Hindi senti lexicon generated by the proposed bootstrap approach contains polar words from the four parts of speech. Hence, identification of correct part of speech of the words in the input document, whose polarity has to be determined is the foremost step. We use Hindi POS tagger tool provided by CDAC[1] to tag the input documents with part of speech. The implemented rule based system is depicted in figure 5.

[1]Available at: http://nlp.cdacmumbai.in/tools.html

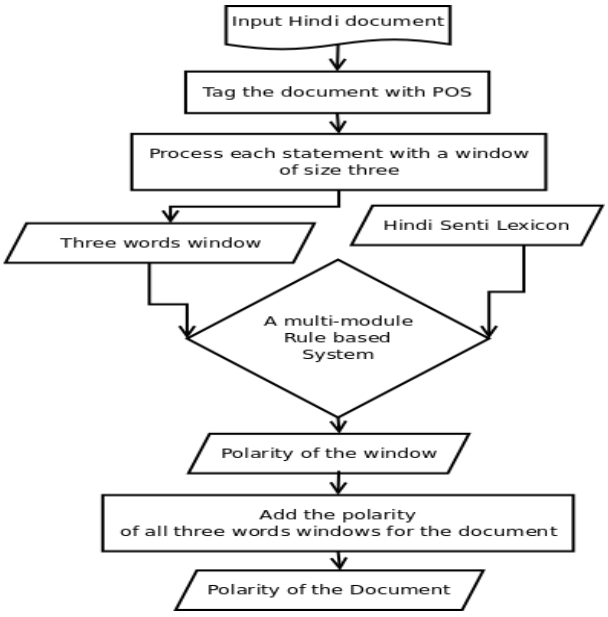

Figure 5: A multi-module sentiment analysis system for Hindi using Hindi senti lexicon

The designed SA system provides polarity of the review document as output. The system processes each sentence/line separately, because sentences are the basic units that determine the polarity of the document. The polarity of a sentence is determined by the analysis of words present in the sentence. The words in a sentence are observed in a window of 3 words to include the effect of several semantic relations among words, for example, the phrase "not so comfortable" has negation before a positive adjective.

The multi-module rule based system decides the polarity of the window of three words such that if there is a word from lexicon with matched part of speech, then the word gets the polarity orientation as given in senti lexicon. The complete polarity of window depends on the neighbor words. For this, the three words window is passed from multiple modules, depending on the part of speech of the words.

Module 1: The presence of negation after a polar word in Hindi reverses the orientation of polarity of the word. This module handles the presence of negation by assigning the reverse polarity of the polar word to the whole window.

Module 2: The presence of two same polarity words in a window enhances the polarity of the posterior polar word. The polarity of the window is determined by adding the individual polarity of each word present in the window.

Module 3: The presence of two opposite polarity words together enhances the polarity of the posterior word (semantic property). If such pattern is observed in the window of three words, this module assigns double polarity of the posterior word to the whole window.

A few instances from all three modules are depicted in figure 6. The ultimate polarity of the document is determined by adding up the polarity of all three words windows of the sentences of the document.

	Semantic Relation	Example Phrase	Polarity of Phrase
Module-1	Polar word + Negation	अभाव (-1) बिल्कुल नहीं	1
Module-2	PosAdj + PosAdj	जबरदस्त (+1) लाजवाब (+1)	2
Module-3	PosAdj + NegNoun	वास्तविक (+1) समस्याएं (-1)	-2
	PosAdj + NegVerb	आसानी (+1) से टूटना (-1)	-2
	PosAdv + NegAdj	अच्छी तरह (+1) अनुपयोगी (-1)	-2
	NegAdj + PosAdj	भयंकर (-1) पसंदीदा (+1)	2
	NegAdj + PosNoun	हद (-1) सिफारिश (+1)	2
	PosAdj + NegAdj	जबर्दस्त (+1) बदसूरत (-1)	-2

Figure 6: Modules of the rule based sentiment analysis system

5 Results of Sentiment Analysis

We validate the efficiency of Hindi senti lexicon generated from the proposed bootstrap approach through the implementation of sentiment analysis (SA) system. We did an extensive validation using two domains: movie and product. Providing sentiment in movie and product domain is a very useful service in current scenario. Its proof is the popularity of Hindi review websites and an umpteen number of Hindi reviews on Web in these two domains.

5.1 Dataset

The movie reviews are manually collected from: *hindi.webdunia.com/entertainment/film/review/*. These reviews are posted by well known critics. If a reviewer has given below 2.5 (below average) star then the review is tagged as negative else the review is tagged as positive. Movie domain dataset contains 100 positive reviews and 100 negative reviews. The average length of the review is 50 words. For product domain, we use the same dataset used by Bakliwal et al. (2012). The dataset contains 350 negative documents and 350 positive documents.

5.2 Accuracy Obtained

Domain	No. of Reviews	Accuracy
Movie	200	**85**
Movie+POS	200	**89.5**
Product	700	**81.4**
Product+POS	700	**85**

Table 2: Average sentiment classification accuracies in percentage

Accuracy is calculated as a fraction of correctly classified documents and total number of documents. The classification accuracies are depicted in table 2 . Presence of polar words in the lexicon from four parts of speech makes it a rich resource for sentiment analysis task. Improvement in accuracy with POS tagging indicates that identification of correct POS tag helps sentiment analysis system. We observe the best accuracy in the movie domain with POS tagging. We compare the results obtained using our sentiment lexicon with the results obtained using graph traversal based subjective lexicon of 8936 words (adjectives and adverbs), presented by Bakliwal et al. (2012) in figure 7. We observe a significant improvement in accuracy in both the domains.

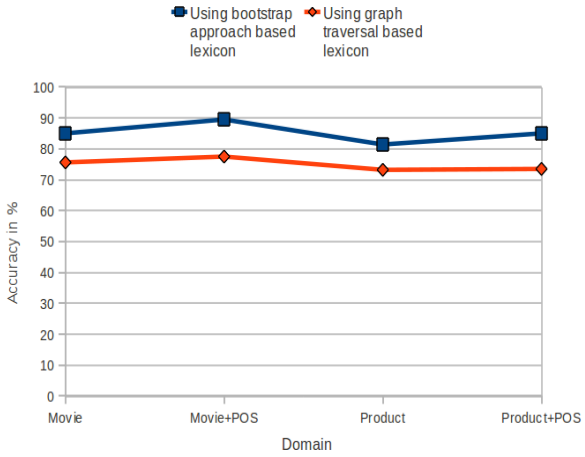

Figure 7: Accuracy obtained with our lexicon and graph traversal based lexicon provided by Bakliwal et al.

6 Related Work

There are many examples, where a lexical resource is used for sentiment analysis in English, but there are a very few instances of sentiment analysis in Hindi. Das and Bandyopadhyay (2009) are the first who report the sentiment analysis in one of the Indian language, that is, Bengali. Since, there is a lack of annotated corpus in Bengali also, their approach was based on sentiment analysis using subjectivity lexicon. Using the lexicon and a few other syntactic features , they achieved a precision of 74.6% and recall of 80.4%.

In case of Hindi, the first work was reported by Joshi et al. (2010). They created a Hindi SentiWordNet using English SentiWordNet and linking between English-HindiWordNet. Using their Hindi SentiWordNet as a lexical resource, they achieved an accuracy of 60.31% for sentiment analysis in Hindi. The reason behind the low accuracy is the distribution of polarity value among the senses of a word in Hindi SentiWordNet, while the testing corpus was not sense tagged. Getting a sense tagged corpora is more expensive than getting the corpora. Joshi et al. reported that a supervised sentiment classification is better than resource based sentiment classification. They got an accuracy of 78.14% using unigram based supervised classification in Hindi.

In our paper, we get an accuracy which is significantly higher than the results reported by Joshi et al. (2010), which indicates that an efficient lexical resource is a better choice than supervised classification in case of lack of corpora. Our work has a close resemblance with the work presented by Bakliwal et al. (2012). The subjectivity lexicon reported by them is used to compare the results (shown in figure 7).

7 Conclusion

In this paper, we present a sentiment lexicon for Hindi, which is a milestone for sentiment analysis in a language lacking in annotated corpus. The proposed bootstrap approach for generation of sentiment lexicon form HindiWordNet is able to extract polar words for all four parts of speech: adjective, noun, adverb and verb.

The senti lexicon obtained from the bootstrap approach is used to build a multi-module rule based sentiment analysis system in Hindi. Multiple modules are embedded to handle the effects of semantic relations among words on polarity of a sentence. The resultant sentiment analysis system is able to produce an average accuracy of 87% for sentiment analysis in the movie and product domain. Besides the betterment of sentiment analysis, the research can be useful for corpora generation, for creating writing aids for authors and in natural language generation.

References

Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon: A lexical resource for hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 141–150, New York, NY, USA. ACM.

Amitava Das and Sivaji Bandyopadhyay. 2009. Subjectivity detection in english and bengali: A crf-based approach. *Proceeding of ICON*.

Aditya Joshi, AR Balamurali, and Pushpak Bhattacharyya. 2010. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 133–140, Stroudsburg, PA, USA. Association for Computational Linguistics.