

A Framework for Sentiment Analysis in Hindi using HSWN

Pooja Pandey

Department of Computer Engineering
PIIT, New Panvel, India

Sharvari Govilkar

Department of Computer Engineering
PIIT, New Panvel, India

ABSTRACT

Due to increase in amount of Hindi content on the web in past years, there are more requirements to perform sentiment analysis for Hindi Language. Sentiment Analysis (SA) is a task which finds orientation of one's opinion in a piece of information with respect to an entity. It deals with analyzing emotions, feelings, and the attitude of a speaker or a writer from a given piece of information. Sentiment Analysis involves capturing of user's behavior, likes and dislikes of an individual from the text. The work of most of the SA system is to identify the sentiments express over an entity, and then classify it into either positive or negative sentiment. Our proposed system for sentiment analysis of Hindi movie review uses HindiSentiWordNet (HSWN) to find the overall sentiment associated with the document; polarity of words in the review are extracted from HSWN and then final aggregated polarity is calculated which can sum as either positive, negative or neutral. Synset replacement algorithm is used to find polarity of those words which don't have polarity associated with it in HSWN. Negation and discourse relations which are mostly present in Hindi movie review are also handled to improve the performance of the system.

General Terms

Sentiment Analysis, HindiSentiWordNet

Keywords

Sentiment Analysis (SA), SentiWordNet, HindiSentiWordNet (HSWN), Polarity, Synset Replacement.

1. INTRODUCTION

Sentiment Analysis is a task under natural language processing which finds orientation of a person opinion or feelings over an entity [1]. It deals with analyzing personal emotions, feelings, attitude and opinion of a speaker or a writer over an object. The primary target of SA is to find the sentiments expressed by person over an information or entity [2].

Hindi is the fourth highest speaking language in the world. The web compared to past years is currently enriched with non English languages too. There exist very few systems which calculate sentiment associated with Hindi text as sentiment analysis is very difficult for Hindi language due different complexity associated with Hindi text. Well annotated standard corpora are still not available for Hindi language. Hindi language lacks availability of efficient resources like parser and tagger which are essential for extracting sentiment. HindiSentiWordNet (HSWN) like well know English SentiWordNet is available but consists of limited numbers of adjectives and adverbs, which still needs to improvement to achieve higher accuracy. There are many

scenarios where same words may be used in multiple contexts and context dependent word mapping is still a difficult task,

error prone and requires manual efforts to find accurate polarity of word.

A framework for performing sentiment analysis on Hindi language is presented in this paper. Related works done and past literature is discussed in section 2. Proposed system is defined which uses HSWN to extract polarity associated with sentiment words in section 3. Finally, section 4 concludes the paper.

2. LITERATURE SURVEY

In this section we cite the relevant past literature of research work done in the field of sentiment analysis for Hindi language.

Opinion Mining System [3] is proposed by authors named as "Hindi Sentiment Orientation System" which is based on Hindi language. Unsupervised approach which based on using dictionary is used to determine the polarity of reviews of user written in Hindi language. Challenges like negation associated in text which reverse the sentiment are also handled. The accuracy of system is evaluated by using 50 sentences of movie reviews and there result showed the accuracy of 65% in finding sentiment associated with text.

Authors proposed HindiSentiWordNet (HSWN) to identify the sentiments associate with Hindi content [4]. Language specific challenges involving negation and discourse which are mostly seen in Hindi text were handled to improve accuracy of the system. Proposed system to find movie reviews in Hindi language achieved approximately 80% accuracy.

Authors have used Semi-Supervised approaches to train a Deep Belief Network on a small percentage of labeled data and assign polarity to unlabelled data [5]. Here sentiment associated with the sentences written in Hindi language is detected using polarity extracted from Deep Belief Network. Here they verify the performance of semi supervised learning with different percentage of labeled data by testing the DBN on validation and the held out test data set. One drawback is that negation rules are not handled by their current unigram count model and they achieved 64% accuracy.

Sentiment analysis for the movie reviews is calculated by the authors [6]. They determine the polarity of sentiments of the person in the review and comments when the sentences occur in documentary level and uses SentiWordNet dictionary to determine the scores of each word present in the comment. They also use rule base and fuzzy logic approach to give the output and achieved the accuracy of 56%.

Author has proposed different approaches like In-language Sentiment Analysis, Resource Based Sentiment Analysis and Machine Translation, to find sentiment associated in Hindi text [1]. HindiSentiWordNet (HSWN) was developed using two lexical resources (English SentiWordNet and English-

Hindi WordNet Linking where Hindi words were matched with corresponding words in English SentiWordNet to extract the polarity associated with the words of Hindi. In-language Sentiment Analysis provided more accuracy than other approaches proposed by the authors.

Authors used WordNet and Breadth First Graph based traversal method to construct the subjectivity lexicon in Hindi for their system to classify polarity associated with the text [7]. A lexicon by using Hindi WordNet involving adverbs and adjectives was created where polarity scores were associated with those words and they also created and annotated corpus of Hindi Product Reviews. For creation of subjective lexicon they also used the relation between synonyms and antonyms where polarity of antonyms is the inverse of the polarity of the synonyms.

3. PROPOSED SYSTEM

To extract sentiment associated with Hindi documents, HindiSentiWordNet (HSWN) will be used which consists of Hindi sentiment words and their associated positive and negative polarity. Here existing HSWN is improved by adding missing sentimental words related to Hindi movie domain. For the input i.e. (Hindi movie review) overall polarity is calculated; which can be positive, negative or it can be neutral, which contribute in finding sentiment of user associated with movie review.

The proposed system comprises two stages:

1. Improving HindiSentiWordNet (HSWN)
2. Sentiment extraction.

First a Hindi movie review dataset will be created, as mostly the data for finding sentiment is usually present on web and bears lot of noises like numbers, one length terms and all which mostly don't contribute in further processing to find sentiment polarity of the input therefore such data will be thoroughly processed to remove unwanted data in the dataset

Our proposed approach performs Sentiment Analysis of Hindi documents using HindiSentiWordNet (HSWN). During the first stage we are improving the existing HSWN with the help of English SentiWordNet, where sentimental words which are not present in the HSWN are translated to English and then searched in English SentiWordNet to retrieve their polarity. In the second stage, sentiment is extracted by finding the overall polarity of the document; which can be positive, negative or neutral. Here during pre-processing tokens are extracted from sentence and spell check is performed. Rules are devised for handling negation and discourse relation which highly influence the sentiments expressed in the document. Finally, overall sentiment orientation of the document is determined by aggregating the polarity values of all the sentimental words in the document.

3.1 Improving HindiSentiWordNet

In this phase existing version of HindiSentiWordNet is improved, as there are limited numbers of adjectives and adverbs i.e. sentiment bearing words present in the existing HSWN so there is need of adding all those missing sentiment bearing words to get more accurate results. HSWN is created using the Hindi WordNet and English SentiWordNet (SWN) by matching the words having same synset IDs. While HSWN was created for Hindi language, it was considered that all synonyms of a particular sentimental word have the same polarity while all antonyms would have the reverse polarity of that word. HSWN is improved by adding those missing adjectives, verbs and adverb sentiment bearing words in the

existing HSWN. Such sentimental words which don't have polarity assigned to it in HSWN are extracted from Hindi text using POS tagger [9] and then polarity of the translated word, using Hindi-English Google translation API, is extracted from English SentiWordNet and then the Hindi word with attached polarity are added to HSWN.

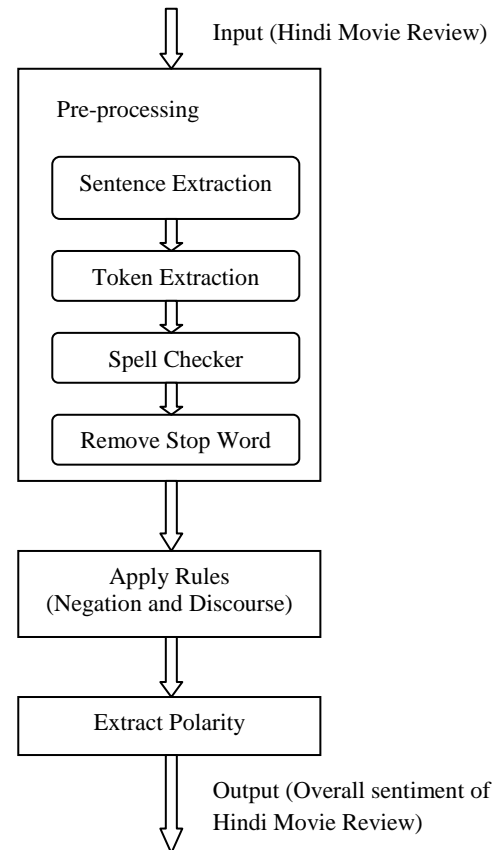


Figure 1. Sentiment Extraction

3.2 Sentiment Extraction

In this stage overall sentiment of the input document will be extracted by using HindiSentiWordNet. This stage consists of three phases:

1. Pre-processing phase
2. Apply Negation and Discourse rules
3. Extracting Polarity

3.2.1 Pre-processing phase:

This phase involves sentence segmentation where paragraph will be segmented into sentences then sentence tokenization will be carried out where tokens will be extracted in the sentence. Spell checking is also performed using online Hindi Spell checker API of Google [10]. Finally stop words will be removed as stop words are those words which don't play much part in further processing to extract sentiment.

3.2.2 Apply Negation and Discourse rules:

This phase involves handling negation and discourse relation in text. The negation operator (नहीं, न, etc.) present in the text mostly inverts the sentimental meaning of the text mostly following the negation words. To handle negation in sentiment analysis first we consider a window of words of a fix size (typically 4 to 6) coming across the negation operator and then reverse the polarity of all the words coming in that

window. Here all the words in the window are reversed by adding (!) symbol to every word in the window, and which is done till either the sentence is completed or an violating expectation or a contrast or conjunction or simply a delimiter is encountered. We have proposed some rules to handle negation on basis of negation operator and sentence structure where inversion may be applied either in forward or in backward direction when negation operator is encountered.

Discourse relations in a text establish a coherent relation between the words in the text; it also links phrases and clauses occurring in a text. Linguistic constructs like conditional, modals, connectives, etc. can alter the associated sentiment present in the text not only at the sentence level but also at the clausal or at phrasal level. Expressions involving words जबकि, हलाकि, लेकिन, etc. are handled by rule created by us, which helps in extracting the actual sentiment associated in the review rather than the miscalculated one.

3.2.3 Extracting Polarity:

In this phase overall sentiment of the document is extracted. Here each token in the document is matched in HSWN to extract its relevant polarity, there may be cases where token polarity is not found for those cases we will use synset replacement algorithm, which replaces the sentiment words with closest meaning word present in the HindiSentiWordNet. If there are still words with no polarity assigned then those words will be saved in separate list so that they can be handled next time to improve HSWN. Finally aggregate polarity of document is found out where we can classify overall polarity as positive, negative or neutral. Synset replacement algorithm [8] replaces the word of which polarity can't be found in the test document with closest word in trained list, having similar sense of which polarity is available in HSWN. Here trained list can be a manually annotated word list or the WordNet.

Following example demonstrate working of our proposed system:

Sample input document:

निर्देशक राजकुमार हिरानी और अभिनेता आमिर खान की जिस फिल्म 'पीके' का सभी को लंबे वक्त से इंतज़ार था, वह आखिरकार रिलीज़ हो गई है। आमिर का काम फिल्म में काबिल-ए-तारीफ है। अनुष्का, सुशांत का काम भी अच्छा है। कुल मिलाकर फिल्म आपका मनोरंजन करेगी।

Pre-processing:

Sentence extraction

Sentence 1: निर्देशक राजकुमार हिरानी और अभिनेता आमिर खान की जिस फिल्म पीके का सभी को लंबे वक्त से इंतज़ार था, आखिरकार रिलीज़ हो गई है।

Sentence 2: आमिर का काम फिल्म में काबिल-ए-तारीफ है।

Sentence 3: अनुष्का, सुशांत का काम भी अच्छा है।

Sentence 4: कुल मिलाकर फिल्म आपका मनोरंजन करेगी।

Stop word removal

Sentence 1: निर्देशक राजकुमार हिरानी और अभिनेता आमिर खान जिस फिल्म पीके सभी लंबे वक्त इंतज़ार, आखिरकार रिलीज़ हो गई ।

Sentence 2: आमिर काम फिल्म काबिल-ए-तारीफ ।

Sentence 3: अनुष्का, सुशांत काम अच्छा ।

Sentence 4: कुल मिलाकर फिल्म आपका मनोरंजन करेगी।

Extracting polarity:

Sentence 1: निर्देशक\0.0 राजकुमार\0.0 हिरानी\0.0 और\0.0 अभिनेता\0.0 आमिर खान\0.0 जिस\0.0 फिल्म\0.0 पीके\0.0 सभी\0.0 लंबे\0.0 वक्त\0.0 इंतज़ार\0.0, आखिरकार\0.0 रिलीज़\0.0 होगई\0.0 ।

Sentence 2: आमिर\0.0 काम\0.375 फिल्म\0.0 काबिल-ए-तारीफ\0.0 ।

After applying synset replacement algorithm आमिर\0.0 काम\0.375 फिल्म\0.0 प्रशंसनीय\0.25।

Sentence 3: अनुष्का\0.0, सुशांत\0.0 काम \0.375 अच्छा\0.45 ।

Sentence 4: कुल मिलाकर\0.0 फिल्म\0.0 आपका\0.0 मनोरंजन\0.23 करेगी\0.0।

Overall polarity: positive.

4. CONCLUSION

Sentiment analysis has lead to determine ones attitude of their speaking or writing through extracting polarity associated with the information. Sentiments can be mined from various sources like texts, tweets, news articles, comments, blogs, social media or from any source of information which is either available online or offline. Sentiment Analysis has been quite popular and has lead to building of better products, understanding user's opinion, executing and managing of business decisions. People rely and make decisions based on reviews and opinions. This research area has provided more importance to the mass opinion instead of word-of-mouth.

This system allows finding sentiment associated with Hindi movie review where overall polarity of the review is classified as positive, negative or neutral using HindiSentiWordNet. To improve the accuracy of finding sentiment synset replacement algorithm is used which helps to find polarity of words which are not available in HSWN.

5. REFERENCES

- [1] Aditya Joshi, Balamurali A, Pushpak Bhattacharyya, "A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study", Proceedings of ICON 2010: 8th International Conference on Natural Language Processing.
- [2] "Sentiment Analysis for Hindi Language", by Piyush Arora 2013.
- [3] Richa Sharma, Shweta Nigam and Rekha Jain, " Polarity Detection Of Movie Reviews In Hindi Language" International Journal on Computational Sciences & Applications (IJCSA) , August 2014.

- [4] Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek "Sentiment Analysis of Hindi Review based on Negation and Discourse relation", International Joint Conference on Natural Language Processing, Nagoya, Japan, October 2013.
- [5] "Sentiment Analysis In Hindi", by Naman Bansal, Umair Z Ahmed, Amitabha Mukherjee.
- [6] Pranali Tumsare, Ashish .S. Sambare and Sachin .R. Jain, "Opinion Mining In Natural Language Processing Using Sentiwordnet and Fuzzy " ,June 2014.
- [7] "Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification " Akshat Bakliwal, Piyush Arora, Vasudeva Varma, 2012.
- [8] Balamurali, A. R., Aditya Joshi, and Pushpak Bhattacharyya, "Robust sense-based sentiment classification," Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, 2011
- [9] <http://text-processing.com/> / accessed on February 10, 2015.
- [10] <https://code.google.com/p/google-api-spelling-java/> /accessed on February 11, 2015.