Sentiment Classification in Hindi by using a HindiSentiWordNet (HSWN)

Pooja Pandey*, Sharvari Govilkar** and Sagar Kulkarni***
*Department of Computer Engg. PIIT, New Panvel, India
pooja.clx@gmail.com
**Department of Computer Engg. PIIT, New Panvel, India
sgovilkar@mes.ac.in
***Department of Computer Engg. PIIT, New Panvel, India
skulkarni@mes.ac.in

Abstract: Sentiment Analysis is a natural language processing task that deals with finding orientation of opinion with respect to a given topic. It deals with analyzing emotions, feelings, and the attitude of a speaker or a writer from a given piece of text or any other form of media. The target of sentiment classification system is to find opinions, identify the sentiments they express, and then classify them according to their polarity. The proposed system for sentiment classification of Hindi documents uses Hindi SentiWordNet (HSWN) to find the overall polarity of the Hindi text document where the final aggregated polarity calculated by system can be positive, negative or neutral. Existing HSWN is enhanced by adding more number of sentiment bearing words. The proposed System also handles negation and discourse relations which influence sentiment associated with a given input.

Keywords: Sentiment Analysis (SA), SentiWordNet, HindiSentiWordNet (HSWN), Polarity, Synset Replacement., Natural language processing (NLP).

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].

Sentiment analysis helps to find sentiment associated with the given input which can be in the form of single line or paragraph or a full document about a given subject. SWN consists of words present in specific language with its associated polarity. For the given input overall polarity or sentiment can be calculated by extracting and aggregating polarity of each sentiment word in the input.

There are different classification levels in SA: document-level, sentence-level and aspect-level. Document-level SA aims to classify an opinion of the whole document as expressing a positive or negative sentiment. Sentence-level SA aims to classify sentiment expressed in each sentence which involves identifying whether sentence is subjective or objective. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities which is done by identifying the entities and their aspects.

The paper presents, sentiment analysis system in Hindi language where overall sentiment is classified as positive or negative. In section 2, proposed system is discussed in detail. Working of system is mentioned in detail in section 3. Section 4 explores accuracy obtained by system. Finally, paper is concluded in section 5.

Related Work

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

Namita mittal et al [1] developed an efficient approach based on negation and discourse relation to identifying the sentiments from Hindi content . They developed an annotated corpus for Hindi language and improve the existing Hindi SentiWordNet (HSWN) by incorporating more opinion words into it.

Aditya Joshi and Pushpak Bhattacharyya [2] proposed a fallback strategy for Hindi language. Authors proposes use of, In-language Sentiment Analysis, Machine Translation and Resource Based Sentiment Analysis to find sentiment in Hindi text. Hindi SentiWordNet (HSWN) was developed using two lexical resources (English SentiWordNet and English-Hindi WordNet Linking .78.14% accuracy was obtained using SVM classifier for in-language sentiment analysis.

Akshat Bakliwal and Piyush [6] present a method of building a subjective lexicon for Hindi. Authors discussed a method of building a subjective lexicon for Hindi. Using WordNet and Breadth First Graph traversal method, they construct the subjectivity lexicon. Main contribution of their work is developing a lexicon of adjectives and adverbs with polarity scores using Hindi WordNet and developing an annotated corpora of Hindi Product Re-views. The limitation of this system is that algorithm does not perform Word Sense Disambiguation. The proposed algorithm achieved ~79% accuracy on classification of reviews and 70.4% agreement with human annotators.

Rekha Jain [5] proposed a Hindi language opinion mining system. In this paper a Hindi language based Opinion Mining System is proposed named as "Hindi Sentiment Orientation System" based on an unsupervised dictionary approach that determine the polarity of user reviews in Hindi language. Negation is also handled in the proposed system. The experiments have been performed by using 50 sentences of movie reviews and achieved the accuracy of 65%.

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. For the input overall polarity is calculated; which can be positive, negative or neutral.

The proposed system consists of two stages:

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

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 stop words are removed. 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.

Improving HindiSentiWordNet

In this phase existing version of HindiSentiWordNet is improved, as it consists of limited numbers of words. 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.

HSWN is improved in two phases by mixing the process of automatic and manual updating of the existing HSWN.

In the first phase all the words which are tagged as adjectives, adverbs and verbs are extracted from English SentiWordNet and then converted to Hindi using BING translator [4]. Now these converted words with attached polarity are added into existing HSWN if they do not already exist in HSWN, to create new updated HSWN named as Improved Hindi SentiWordNet "Improved-HSWN".

In the second phase a dataset is created which consists of Hindi reviews extracted from web. From this dataset sentiment words are extracted which are not present in Improved-HSWN. Around 700 such words where extracted and polarity was manually assigned and then added to Improved-HSWN.

Finally an improved HSWN is available which consists of 28703 words as compared to 11941 words present in HSWN provided by IIT Bombay [3]. Improving existing SentiWordNet is important part as more the count of sentiment words with polarity then there is better chance for system to find accurate sentiment associated with the input text.

Sentiment Extraction

In this stage overall sentiment of the input document will be extracted by using Improved-HSWN. By using Improved-HSWN polarity of words present in the document are extracted one by one and aggregated together to calculate the overall polarity associated with the document. The overall polarity is then classified as positive, negative or neutral to specify the sentiment associated with the document.

This stage consists of three sub modules:

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

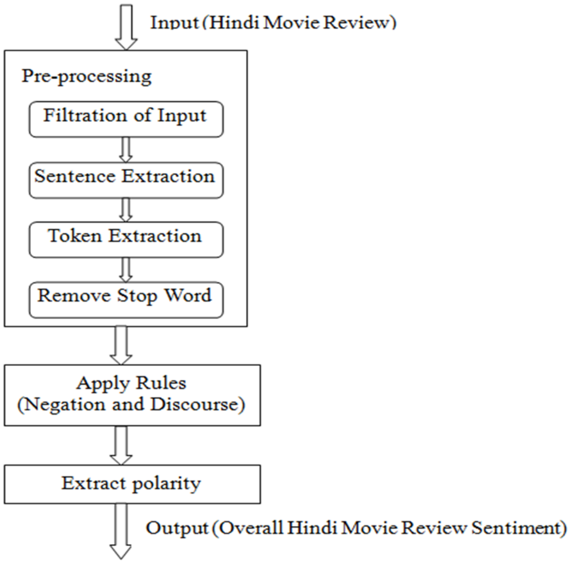


Figure 1. Proposed System

```
Algorithm for extracting polarity from the text
Step 1: For each token in the document.
Step 2: Check If (word is present in Improved-HSWN)
        Then Retrieve polarity (POL) from Improved-HSWN and go to Step 5
        Else using synset replacement algorithm replaces the sentiment words with closest  meaning word present in the Improved-HSWN.
Step 4: If (no polarity assigned to word)
        Then fetch next token and go to Step 2
Step 5: If (word is negated)
        Then word POL= - (POL);
        Else word POL remains the same
Step 5: If (word is marked for discourse)
        Then word POL= weight*(POL);
        Else word POL= (POL);
        End For Loop and go to Step 6 when all tokens are processed.
Step 6: Compute the aggregate polarity of the document (doc POL) by adding the polarities values of the entire tokens.
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
```

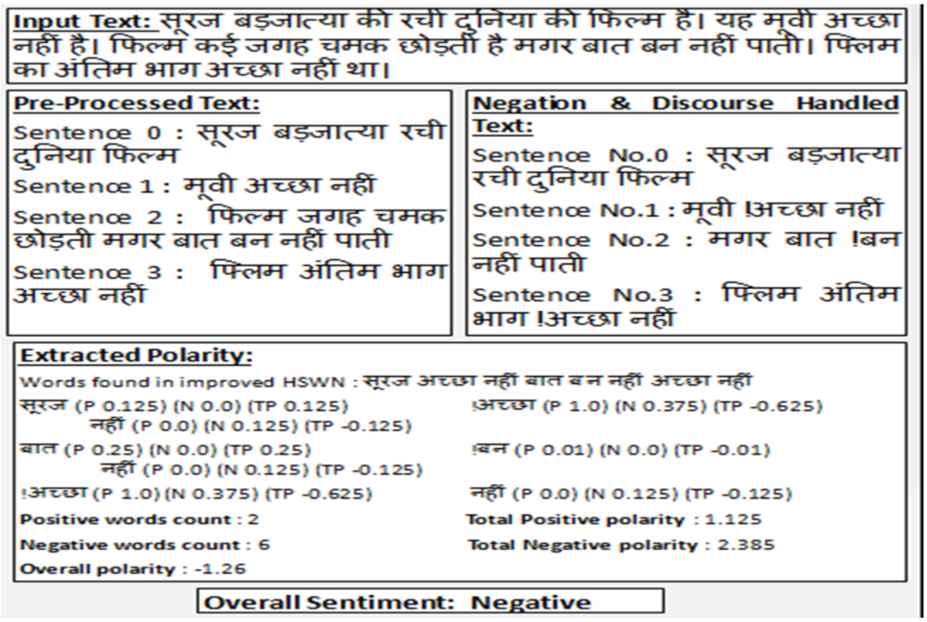
Working of System

The input to the system is a single text document in Hindi. The system accepts a .txt file as input or user can write there reviews in the provided text area using any offline or online Hindi input tool. Working model of system is shown in Figure 2. The input to the system is a single text document in Hindi. The system accepts a .txt file as input or user can write there reviews in the provided text area using any offline or online Hindi input tool.

Improving HSWN

To improve the existing HSWN provided by IIT Bombay first complete existing English SentiWordNet is translated into Hindi using Bing translator and if translated word doesn't exist in existing HSWN then it's added to the existing HSWN.

As our target domain to extract sentiment is of Hindi Movie reviews so to improve existing HSWN by adding more movie domain related words into it we have collected overall 235 Hindi movie reviews from different online movie reviewing sites. Batch processing is performed where all unwanted characters and symbols and stop word are removed from it initially. Then words which are not found in Improved- HSWN are assigned polarity manually and added in the Improved-HSWN.



Input Text: सूरज बड़जात्या की रची दुनिया की फिल्म है। यह मूवी अच्छा नहीं है। फिल्म कई जगह चमक छोड़ती है मगर बात बन नहीं पाती। फ्लिम का अंतिम भाग अच्छा नहीं था।

Pre-Processed Text:

Sentence 0 : सूरज बड़जात्या रची दुनिया फिल्म

Sentence 1 : मूवी अच्छा नहीं

Sentence 2 : फिल्म जगह चमक छोड़ती मगर बात बन नहीं पाती

Sentence 3 : फ्लिम अंतिम भाग अच्छा नहीं

Negation & Discourse Handled Text:

Sentence No.0 : सूरज बड़जात्या रची दुनिया फिल्म

Sentence No.1 : मूवी !अच्छा नहीं

Sentence No.2 : मगर बात !बन नहीं पाती

Sentence No.3 : फ्लिम अंतिम भाग !अच्छा नहीं

Extracted Polarity:

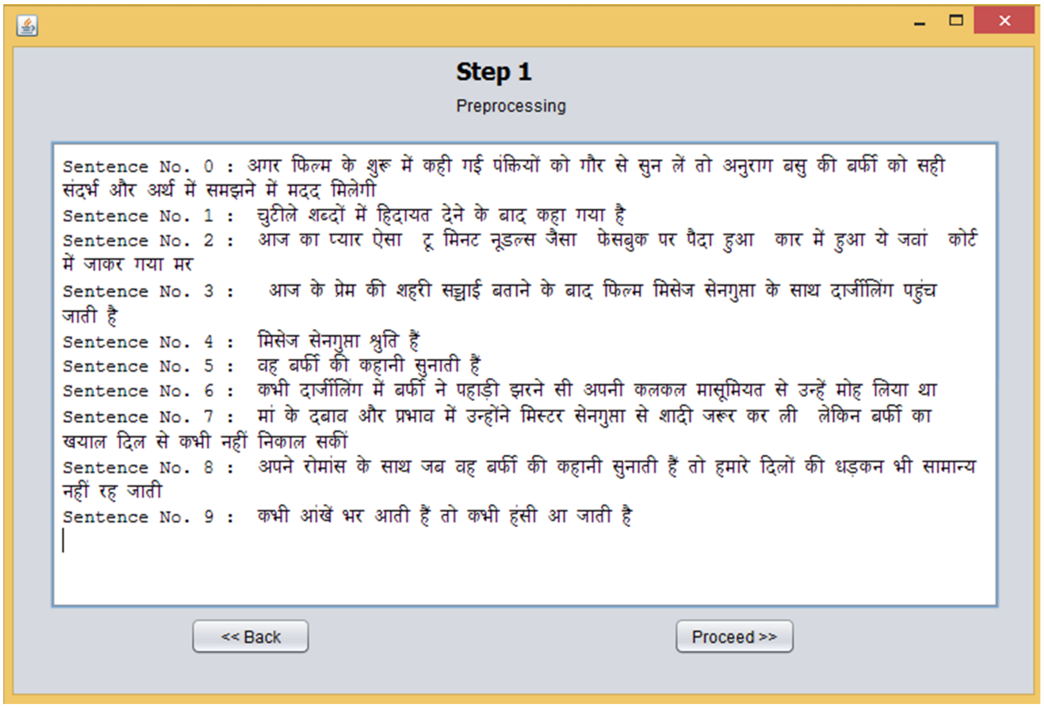
```
Words found in improved HSWN : सूरज अच्छा नहीं बात ब न नहीं अच्छा नहीं
सूरज (P 0.125) (N 0.0) (TP 0.125)                 !अच्छा (P 1.0) (N 0.375) (TP -0.625)
        नहीं (P 0.0) (N 0.125) (TP -0.125)
बात (P 0.25) (N 0.0) (TP 0.25)                    !बन (P 0.01) (N 0.0) (TP -0.01)
         नहीं (P 0.0) (N 0.125) (TP -0.125)
!अच्छा (P 1.0) (N 0.375) (TP -0.625)               नहीं (P 0.0) (N 0.125) (TP -0.125)
Positive words count : 2                          Total Positive polarity : 1.125
Negative words count : 6                          Total Negative polarity : 2.385
Overall polarity : -1.26
```

Overall Sentiment: Negative

Figure 2. Working Model

Sentiment Extraction

Here in the first step input text is preprocessed to remove unwanted characters symbols and stop words from the input text. Here in filtration of input text non Devanagari Unicode characters are removed from the input text. Also symbols and punctuation and numbers expect "," and "|" are removed from the input text. Here individuals sentence are separated and stored in a list from the filtered input text. Each sentence is given a Sentence ID to it as shown in Figure 3.



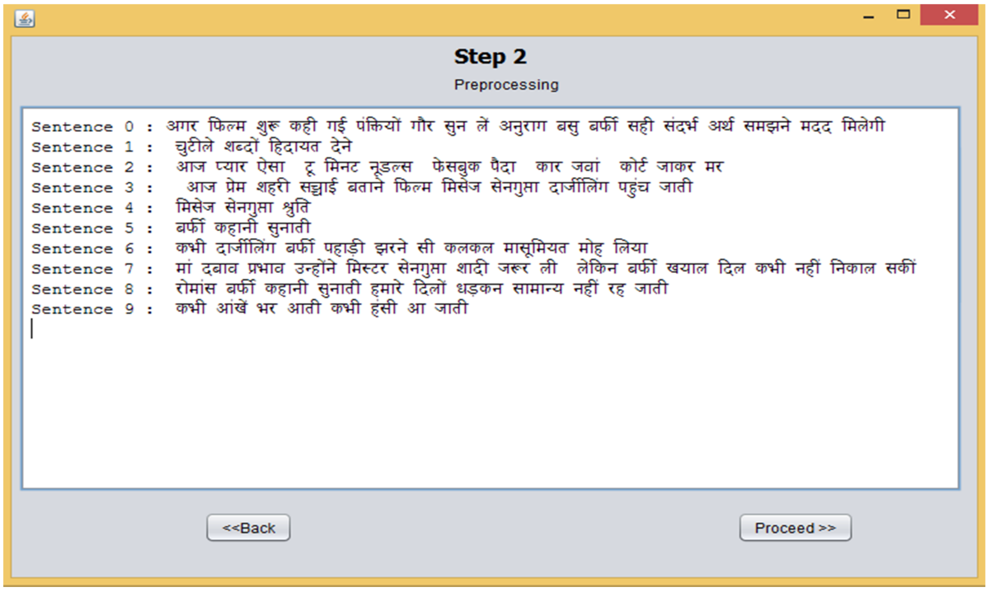
Step 1

Preprocessing

```
Sentence No. 0 : अगर फिल्म के शुरू में कही गई पंक्तियों को गौर से सुन लें तो अनुराग बसु की बर्फी को सही
संदर्भ और अर्थ में समझने में मदद मिलेगी
Sentence No. 1 :  चुटीले शब्दों में हिदायत देने के बाद कहा गया है
Sentence No. 2 :  आज का प्यार ऐसा  टू मिनट नूडल्स जैसा  फेसबुक पर पैदा हुआ  कार में हुआ ये जवां  कोर्ट
में जाकर गया मर
Sentence No. 3 :   आज के प्रेम की शहरी सच्चाई बताने के बाद फिल्म मिसेज सेनगुप्ता के साथ दार्जीलिंग पहुंच
जाती है
Sentence No. 4 :  मिसेज सेनगुप्ता श्रुति हैं
Sentence No. 5 :  वह बर्फी की कहानी सुनाती हैं
Sentence No. 6 :  कभी दार्जीलिंग में बर्फी ने पहाड़ी झरने सी अपनी कलकल मासूमियत से उन्हें मोह लिया था
Sentence No. 7 :  मां के दबाव और प्रभाव में उन्होंने मिस्टर सेनगुप्ता से शादी जरूर कर ली  लेकिन बर्फी का
खयाल दिल से कभी नहीं निकाल सकीं
Sentence No. 8 :  अपने रोमांस के साथ जब वह बर्फी की कहानी सुनाती हैं तो हमारे दिलों की धड़कन भी सामान्य
नहीं रह जाती
Sentence No. 9 :  कभी आंखें भर आती हैं तो कभी हंसी आ जाती है
```

Figure 3. Filtration and extraction of sentence from input text

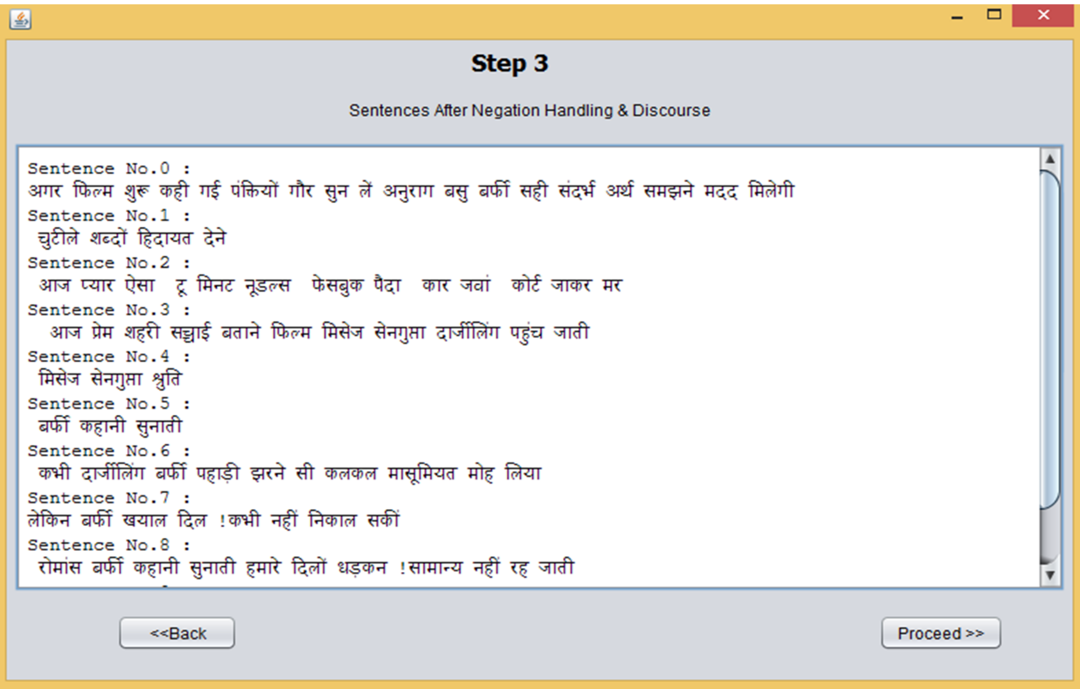
Stop words which don't provide any relevance to extract overall sentiment are removed from the tokenized input text.



```
Sentence 0 :  अगर फिल्म शुरू कही गई पंक्तियों गौर सुन लें अनुराग बसु बर्फी सही संदर्भ अर्थ समझने मदद मिलेगी
Sentence 1 :  चुटीले शब्दों हिदायत देने
Sentence 2 :  आज प्यार ऐसा  टू मिनट नूडल्स  फेसबुक पैदा  कार जवां  कोर्ट जाकर मर
Sentence 3 :   आज प्रेम शहरी सच्चाई बताने फिल्म मिसेज सेनगुप्ता दार्जीलिंग पहुंच जाती
Sentence 4 :  मिसेज सेनगुप्ता श्रुति
Sentence 5 :  बर्फी कहानी सुनाती
Sentence 6 :  कभी दार्जीलिंग बर्फी पहाड़ी झरने सी कलकल मासूमियत मोह लिया
Sentence 7 :  मां दबाव प्रभाव उन्होंने मिस्टर सेनगुप्ता शादी जरूर ली  लेकिन बर्फी खयाल दिल कभी नहीं निकाल सकीं
Sentence 8 :  रोमांस बर्फी कहानी सुनाती हमारे दिलों धड़कन सामान्य नहीं रह जाती
Sentence 9 :  कभी आंखें भर आती कभी हंसी आ जाती
```

Figure 4. Stop words removed

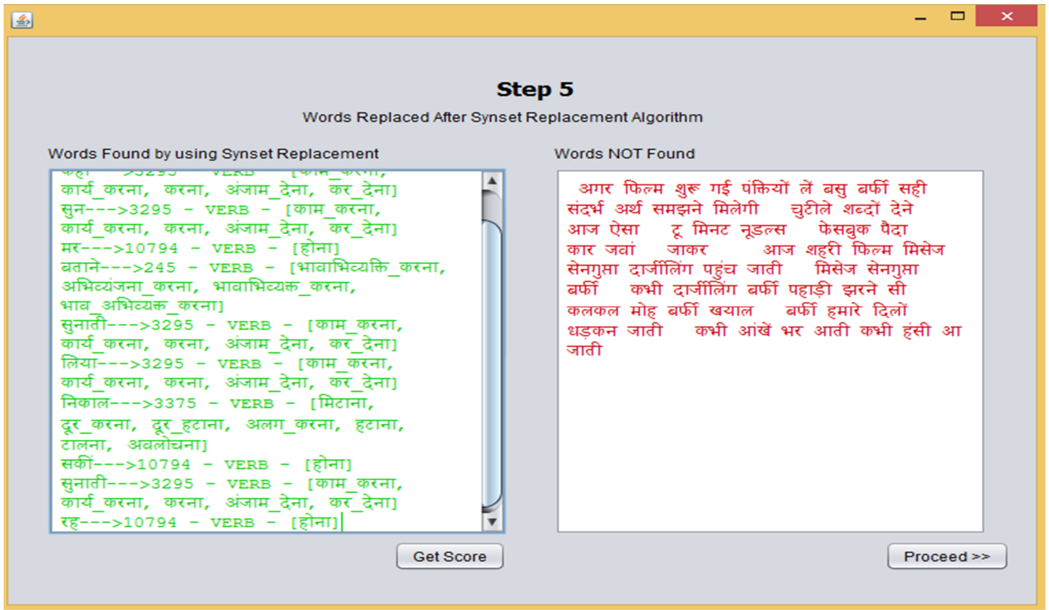
Words which are candidate to be negated are assigned “!” mark next to it. Such negated words polarity will be reversed in further stages.



```
Sentence No.0 :
अगर फिल्म शुरू कही गई पंक्तियों गौर सुन लें अनुराग बसु बर्फी सही संदर्भ अर्थ समझने मदद मिलेगी
Sentence No.1 :
 चुटीले शब्दों हिदायत देने
Sentence No.2 :
 आज प्यार ऐसा  टू मिनट नूडल्स  फेसबुक पैदा  कार जवां  कोर्ट जाकर मर
Sentence No.3 :
  आज प्रेम शहरी सच्चाई बताने फिल्म मिसेज सेनगुप्ता दार्जीलिंग पहुंच जाती
Sentence No.4 :
 मिसेज सेनगुप्ता श्रुति
Sentence No.5 :
 बर्फी कहानी सुनाती
Sentence No.6 :
 कभी दार्जीलिंग बर्फी पहाड़ी झरने सी कलकल मासूमियत मोह लिया
Sentence No.7 :
लेकिन बर्फी खयाल दिल !कभी नहीं निकाल सकीं
Sentence No.8 :
 रोमांस बर्फी कहानी सुनाती हमारे दिलों धड़कन !सामान्य नहीं रह जाती
```

Figure 5. Negation and discourse handling

Words of which polarity is not found in Improved-HWSN are then processed using sense based synset replacement algorithm to find polarity of word by finding the word having polarity with the same synset ID as the input word.



Step 5

Words Replaced After Synset Replacement Algorithm

Words Found by using Synset Replacement

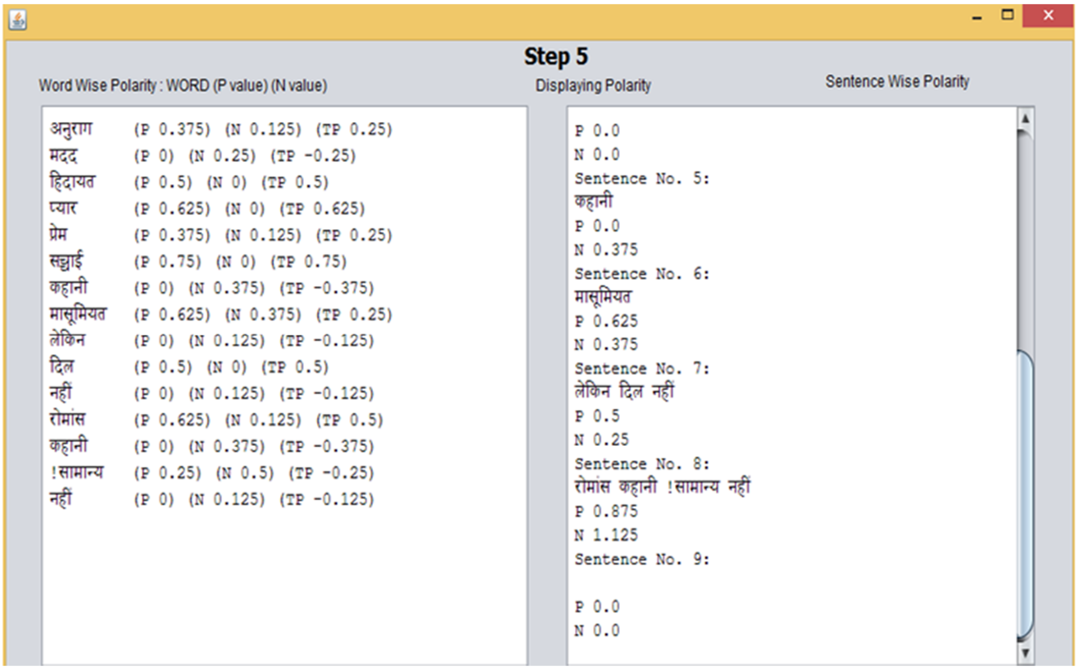
```
कहा--->3295 - VERB - [काम_करना,
कार्य_करना, करना, अंजाम_देना, कर_देना]
सुन--->3295 - VERB - [काम_करना,
कार्य_करना, करना, अंजाम_देना, कर_देना]
मर--->10794 - VERB - [होना]
बताने--->245 - VERB - [भावाभिव्यक्ति_करना,
अभिव्यंजना_करना, भावाभिव्यक्त_करना,
भाव_अभिव्यक्त_करना]
सुनाती--->3295 - VERB - [काम_करना,
कार्य_करना, करना, अंजाम_देना, कर_देना]
लिया--->3295 - VERB - [काम_करना,
कार्य_करना, करना, अंजाम_देना, कर_देना]
निकाल--->3375 - VERB - [मिटाना,
दूर_करना, दूर_हटाना, अलग_करना, हटाना,
टालना, अवलोचना]
सकीं--->10794 - VERB - [होना]
सुनाती--->3295 - VERB - [काम_करना,
कार्य_करना, करना, अंजाम_देना, कर_देना]
रह--->10794 - VERB - [होना]
```

Words NOT Found

```
 अगर फिल्म शुरू गई पंक्तियों लें बसु बर्फी सही
संदर्भ अर्थ समझने मिलेगी   चुटीले शब्दों देने
आज ऐसा   टू मिनट नूडल्स   फेसबुक पैदा
कार जवां   जाकर     आज शहरी फिल्म मिसेज
सेनगुप्ता दार्जीलिंग पहुंच जाती   मिसेज सेनगुप्ता
बर्फी   कभी दार्जीलिंग बर्फी पहाड़ी झरने सी
कलकल मोह बर्फी खयाल   बर्फी हमारे दिलों
धड़कन जाती   कभी आंखें भर आती कभी हंसी आ
जाती
```

Figure 6. Synset replacement

Words are assigned polarity to it by fetching of polarity from the Improved-HSWN. Word polarity value can be positive and negative like a word “X” is assigned polarity value as positive “0.0” and negative “0.1”.



Step 5

Word Wise Polarity : WORD (P value) (N value)

```
अनुराग     (P 0.375) (N 0.125) (TP 0.25)
मदद       (P 0) (N 0.25) (TP -0.25)
हिदायत     (P 0.5) (N 0) (TP 0.5)
प्यार       (P 0.625) (N 0) (TP 0.625)
प्रेम        (P 0.375) (N 0.125) (TP 0.25)
सच्चाई     (P 0.75) (N 0) (TP 0.75)
कहानी     (P 0) (N 0.375) (TP -0.375)
मासूमियत  (P 0.625) (N 0.375) (TP 0.25)
लेकिन     (P 0) (N 0.125) (TP -0.125)
दिल       (P 0.5) (N 0) (TP 0.5)
नहीं       (P 0) (N 0.125) (TP -0.125)
रोमांस     (P 0.625) (N 0.125) (TP 0.5)
कहानी     (P 0) (N 0.375) (TP -0.375)
!सामान्य   (P 0.25) (N 0.5) (TP -0.25)
नहीं       (P 0) (N 0.125) (TP -0.125)
```

Displaying Polarity

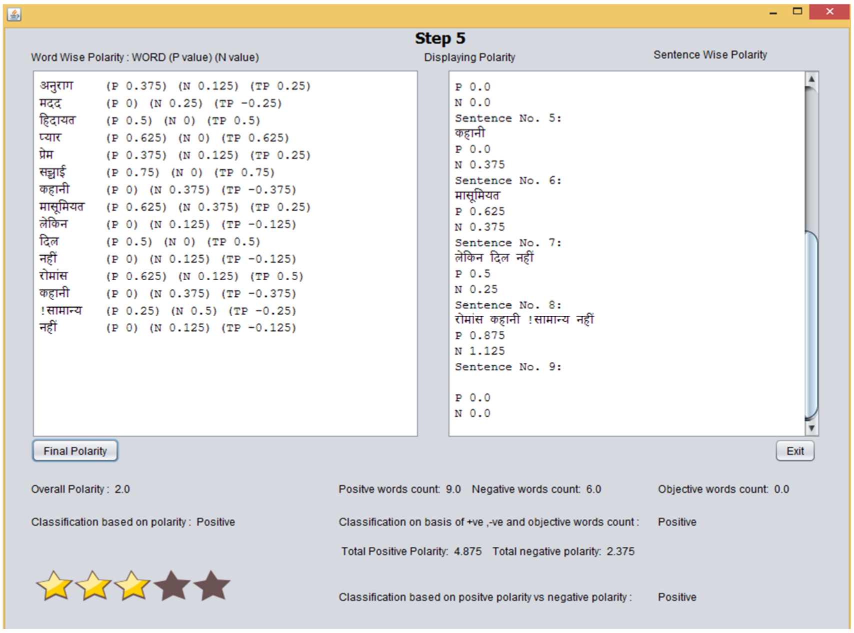
Sentence Wise Polarity

```
P 0.0
N 0.0
Sentence No. 5:
कहानी
P 0.0
N 0.375
Sentence No. 6:
मासूमियत
P 0.625
N 0.375
Sentence No. 7:
लेकिन दिल नहीं
P 0.5
N 0.25
Sentence No. 8:
रोमांस कहानी !सामान्य नहीं
P 0.875
N 1.125
Sentence No. 9:

P 0.0
N 0.0
```

Figure 7. Polarity of each token in the input text

Finally polarities of all words are aggregated to find total overall polarity of the input text and sentiment associated with the input text is presented by classifying the polarity value as positive negative or neutral.



Step 5

Word Wise Polarity : WORD (P value) (N value)

```
अनुराग     (P 0.375) (N 0.125) (TP 0.25)
मदद       (P 0) (N 0.25) (TP -0.25)
हिदायत     (P 0.5) (N 0) (TP 0.5)
प्यार      (P 0.625) (N 0) (TP 0.625)
प्रेम       (P 0.375) (N 0.125) (TP 0.25)
सच्चाई     (P 0.75) (N 0) (TP 0.75)
कहानी      (P 0) (N 0.375) (TP -0.375)
मासूमियत   (P 0.625) (N 0.375) (TP 0.25)
लेकिन      (P 0) (N 0.125) (TP -0.125)
दिल        (P 0.5) (N 0) (TP 0.5)
नहीं        (P 0) (N 0.125) (TP -0.125)
रोमांस      (P 0.625) (N 0.125) (TP 0.5)
कहानी      (P 0) (N 0.375) (TP -0.375)
!सामान्य    (P 0.25) (N 0.5) (TP -0.25)
नहीं        (P 0) (N 0.125) (TP -0.125)
```

Displaying Polarity

Sentence Wise Polarity

```
P 0.0
N 0.0
Sentence No. 5:
कहानी
P 0.0
N 0.375
Sentence No. 6:
मासूमियत
P 0.625
N 0.375
Sentence No. 7:
लेकिन दिल नहीं
P 0.5
N 0.25
Sentence No. 8:
रोमांस कहानी !सामान्य नहीं
P 0.875
N 1.125
Sentence No. 9:

P 0.0
N 0.0
```

Overall Polarity : 2.0

Positve words count: 9.0 Negative words count: 6.0

Objective words count: 0.0

Classification based on polarity : Positive

Classification on basis of +ve ,-ve and objective words count : Positive

Total Positive Polarity: 4.875 Total negative polarity: 2.375

☆☆☆★★

Classification based on positve polarity vs negative polarity : Positive

Figure 8. Sentiment classification of input text

Performance Analysis

The system is evaluated to check whether the output generated is correct or not. The calculations are performed to check whether the given input is classified correctly or not. A given input can be classified as positive, negative or neutral by the system. Performance of system is evaluated by analyzing how much accurately system extract sentiment associated with the given input.

Accuracy is the measure used here is to evaluate the system. Accuracy of the system is calculated by using following formula: Accuracy = ((Total number of documents which are classified correctly) / (Total number of document present))*100.

To calculate accuracy of system overall 90 Hindi documents having opinions or sentiments are considered of varying domain like news, people, place and technology. Also Hindi tweets are processed to extract sentiment. Out of this 90 documents around 73 documents where correctly classified.

Comparison of accuracy for 90 Hindi documents between original HSWN and Improved HSWN is show in below figure.

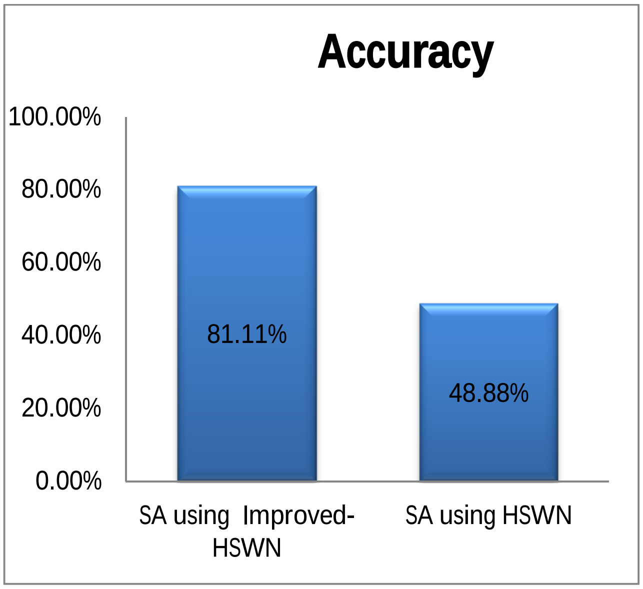


Figure 9. Comparison of system accuracy using HSWN vs. Improved-HSWN for movie reviews

Conclusion

Sentiment analysis has lead to determine the attitude or inclination of a communicator through the contextual polarity of their speaking or writing. Sentiments can be mined from texts, tweets, blogs, social media, news articles, comments or from any source of information.

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, with the system in their daily spoken natural language.

In future system can be updated to add more sentimental words to improve accuracy; also other sentiment analysis challenges like sarcasms and implicit opinion can be handled. Also tools like word sense disambiguation can be used which can help in correct extraction of word polarity based on word actual sense. The work can be extended to support other regional languages like Marathi, Kannada, Guajarati, Manipur etc.

Acknowledgment

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks to the head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

References

[1] 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.

[2] Joshi, Aditya, A. R. Balamurali, and Pushpak Bhattacharyya, "A fall-back strategy for sentiment analysis in Hindi: a case study," Proceedings of the 8th ICON (2010).

[3] http://www.cfilt.iitb.ac.in/Sentiment_Analysis_Resources.html last accessed on September 13, 2015 at 5.35 p.m.

[4] http://www.microsoft.com/en-us/translator/translatorapi.aspx last accessed on October 6, 2015 at 4.02 p.m .

[5] Richa Sharma, Shweta Nigam and Rekha Jain, "Polarity Detection Of Movie Reviews In Hindi Language" International Journal on Computational Sciences & Applications (IJCSA), August 2014.

[6] Akshat Bakliwal, Piyush Arora, Vasudeva Varma, " Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification", Search and Information Extraction Lab, LTRC,2012.