Sentiment Analysis on Hindi News Articles

Prof. Omprakash Yadav*, Rahul Patel[1], Yash Shah[2], Saneesha Talim[3]

**Professor, Department of Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India*
[1,2,3]B. E student, Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India

---***---

Abstract – *Due to the recent boom in the number of Internet users in India. It led to the increase in different types of data available on the internet, resulting in availability of information in various regional as well as national language. In India, Hindi Language Interfaces or technology are absent to a great extent. With that, there was requirement for such system. Our system is based on the application of sentiment analysis for Hindi news articles using natural language processing technique and computational linguistics. Sentiment analysis helps us understand if a piece of information is positive, negative, or neutral. Sentiment analysis helps understand or identify the attitude and emotional state of an individual, through its written piece. The feature of this work is that it contains the dictionary of all negative and positive words which help to analyse the polarity more precisely.*

***Key Words*: Sentiment Analysis, Natural Language Processing, Computational Linguistics, Polarity.**

1.INTRODUCTION

Hindi is fourth highest speaking language in the world. It is spoken nearly by 430 million people as primary language and 125 million people as a secondary language. Various websites, blogs and tweets have started supporting Hindi language and some of them use Hindi as a first language as well. We have found out that, in the real world, people are often comfortable in speaking, understanding, and writing in their native language. Hindi is India's national language and is spoken in almost every house. But with the sudden increase in web content and internet user, most of the information available on the internet is in English. As English is not dominant as Hindi is in India.

Many organization, commercial and social business provide Hindi content on web for the better understanding among the users. Such circumstances have motivated us to carry out research in sentiment analysis on the same. But building such system is a very difficult task due to limited resources available on the internet about Hindi Language as compared to English. One of the main challenging tasks to carry out sentiment analysis on Hindi language is that a particular word can be used in multiple contexts.

Sentiment Analysis helps in understanding and categorizing emotions within a given text. These emotions can be positive, negative, or neutral. A written piece can be conceptual based, facts or sentiments. To understand the difference between them is crucial. If we treat all of them in the same way, the result will not be prompt and accurate. So, it is mandatory that we first understand the context and identify to which category it belongs to. It concentrates organizing at many levels of various natures. Today, it has a wide application, especially in the fields of marketing, customer services. It concentrates on sorting the content as per the required subject. Application of this are present across different domain like sociology, psychology, etc. Such algorithms also hold a dominant part in feedback and recommendation systems. Major companies like Facebook and Uber also uses this technology for Intent Analysis and for better Customer Relationship Model. Due to the various structures, it is also referred as Opinion Mining.

In our approach, we are scraping a news article from a Hindi news website and then extracting sentiments from the body of that particular article. This paper will discuss architecture. algorithm and datasets used to carry out the function.

2. Existing Work

A lot of development has been made in the area of sentimental analysis on English Language but the work in Hindi is very limited. Nevertheless, there is been a considerable progress in past few year. Some of them were:

A. Joshi, B. A. R, and P. Bhattacharyya [1] suggested a strategy for sentiment analysis of Hindi language. In this model,3 approaches were used- In-language translation, machine translation, resource-based sentiment analysis. In this paper, each of the methods have their own advantages. The first approach involves creating corpora for Hindi Movie Domain and developing a classifier to classify aa new Hindi Text. The Second process includes translating a dataset from English to Hindi further by applying the previous classifier. In the third method, a lexical database called as HindiSentiWordNet is used to classify a text.

K, Bandyopadhyay [2] anticipated a verb-based method for Sentiment analysis for Manipuri language. In this model an unsupervised learning approach called CRF (Conditional Random Field) is used. They also proposed a model for other regional language.

N. Mittal, B. Agarwal, G. Chouhan [3] studied on the Hindi language content. In this paper, it is examined that by appropriate handling of negation and discourse relation it may improve results in comparison to other existing methods.

3. Proposed Work

The proposed work is similar to [1] using a dataset of influential words with both positive and negative polarity with additional inclusion of rare words and checking the overall polarity followed by stop words removal. Serval databases are given as input to the system is given at different points. Stopwords [9] is used in Article summarization process and positive, negative words [8] are used for negation handling and polarity calculation.

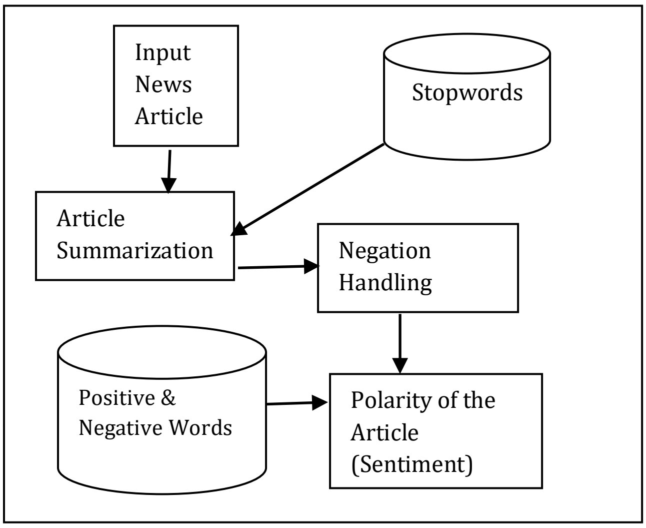


Fig -1: Block Diagram for proposed system

3.1 Algorithm

```
1. Input Article URL
2. Scrape Data from the Webpage like article text and
   heading
3. Perform Pre-processing techniques such as
   Segmentation, Tokenization and Stop-words
   Removal for Text Summary
4. Compare all the words with positive and negative
   words list
   If match found
   Increment the corresponding counters
   Else
   Continue
5. For remaining words in the sentence
   Check for Negation property
   If found
   Allocate reverse polarity
   Else
   Continue
6. If other words found repeat step 4
   Else
   Go to Step 7
7. Compare Positive and Negative pointer
8. Display the polarity.
```

3.2 Article Summarization

Article Summarization is a process analysing a document or a piece of written information to extract all the crucial information which can represent the whole context of the information in just limited numbers of words. The process of text summarization is divided into 3 sub-processes namely, Segmentation, Tokenization and Stop-Words Removal.

3.2.1 Segmentation

In Hindi language sentence endings is defined by "|". But nowadays, Hindi news websites like BBC, AajTak, etc. have started using full stop "." rather than "|". So, segmentation breaks paragraphs into different sentences and every sentence is stored in a list. The list is required for further processing.

3.2.2 Tokenization

In this part, sentences are tokenized by identifying commas and spaces between the words and a separate word list is maintained. The list created by tokenization process is called tokens. This process also helps us to achieve single type of text by removing special characters, commas , etc.A single word is also referred as tokens.

3.2.3 Stop-Words Removal

Most regularly used words are called stopwords. Stop-words are insignificant and do not have any importance. So, these types of words should be removed from the text, otherwise it can affect the polarity values. We observed that every Hindi text covers at most 25-30% or more stop-words. We have developed a dataset using different sources from internet which has huge variety of stop words which can contribute unintentionally in achieving accuracy and better opinion prediction. Example of stop-words are "का"," ही"," हुआ" etc.

3.3 Negation Handling

Since Hindi language is known for its nature of being unstructured. Negation handling for this language can be quite difficult. This stage involves treating negation in text. The negation operator (नहीं, न, etc.) present in the text mostly inverts the meaning of the text which affects the sentiment score or polarity in a critical way. To handle situation first we consider a couple of words of size (4 to 7). Mostly the system looks for the occurrences of "नहीं". On encountering a negation operator assign reverse polarity to all the words that appears in that particular window.

Positive Sentence

यह बिलकुल बुरी खबर नहीं है

Negative Sentence

यह खुशकिस्मती की बात नहीं है

Fig -2: Sample for Negation Handling

3.4 Polarity of Article

Here the article or text is checked for number of matches found in dictionary and compared. The words with positive, negative and neutral opinion are assigned value 1, -1 and 0 respectively. And the majority of the words of same polarity critically affects the sentiment score. The nature of polarity can be identified by the score. If the score is greater than zero then it is identified as positive sentiment. The score less than zero is categorized as negative opinion. If then sentiment score is zero than it is considered as neutral.

4. Experimentation and Results

In this section we discuss about the experiment conducted on Hindi News Sentiment Analysis. For implementing the discussed approach, we have used Python 3 on Google Collaboratory.

4.1 Dataset

As main focus of the algorithm is based on examining singular words. We have gathered several data set across internet to make a unified database for multipurpose approach.

Table -1: Characteristics of Dataset

Words	No of Words
Positive Words	1342
Negative Words	1409
Stop Words	372

4.2 Result

Experiment is conducted on Hindi News Article using static dataset which contained Hindi words from various datasets. Input is given to the system and output is processed in the system by comparing the positive and negative matches with negation handling. For measuring accuracies will be calculating with and without negation for better understanding. The common way for computing these measures is based on the Polarity matrix shown below.

Table -2: Polarity Matrix

Words	Predicted Positives	Predicted Negatives
Positive Words	# of true positive words(tp)	# of false negative words(fn)
Negative Words	# of false positive words(fp)	# of true negative words(tn)

Three evaluation measures are used on the basis of which system performance is computed; these are:

1.Precision: It is the ratio of true positive predicted instances against all positive predicted instances.

Precision= tp/(tp+fp)

2.Recall: It is the ratio of true positive predicted instances against all actual positive instances.

Recall =tp/(tp+fn)

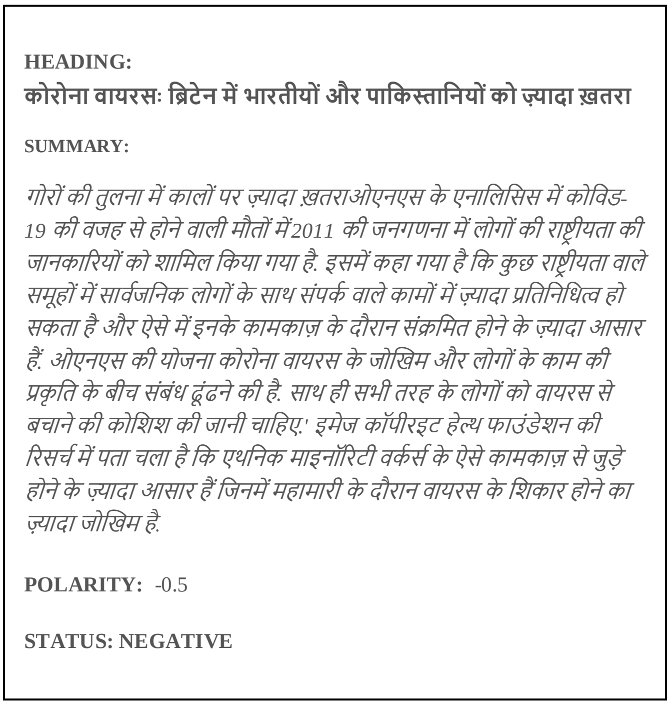
3.Accuracy: It is the ratio of true predicted instances against all predicted instances.

Accuracy=(tp+tn)/(tp+fp+tn+fn)

Table -3: Results

Method	Precision	Recall	Accuracy
With Negation	82.15 %	95.23 %	83.29 %
Without Negation	83.03 %	96.12	81.02%

Refer to Fig-3 for the practical output of the proposed system. We have taken a URL for a Hindi News article from BBC. The image below will show the heading for that respective article, summarized text and polarity of that particular summary.



HEADING:
कोरोना वायरसः ब्रिटेन में भारतीयों और पाकिस्तानियों को ज़्यादा ख़तरा

SUMMARY:

गोरों की तुलना में कालों पर ज़्यादा ख़तराओएनएस के एनालिसिस में कोविड-19 की वजह से होने वाली मौतों में2011 की जनगणना में लोगों की राष्ट्रीयता की जानकारियों को शामिल किया गया है. इसमें कहा गया है कि कुछ राष्ट्रीयता वाले समूहों में सार्वजनिक लोगों के साथ संपर्क वाले कामों में ज़्यादा प्रतिनिधित्व हो सकता है और ऐसे में इनके कामकाज़ के दौरान संक्रमित होने के ज़्यादा आसार हैं. ओएनएस की योजना कोरोना वायरस के जोखिम और लोगों के काम की प्रकृति के बीच संबंध ढूंढने की है. साथ ही सभी तरह के लोगों को वायरस से बचाने की कोशिश की जानी चाहिए.' इमेज कॉपीरइट हेल्थ फाउंडेशन की रिसर्च में पता चला है कि एथनिक माइनॉरिटी वर्कर्स के ऐसे कामकाज़ से जुड़े होने के ज़्यादा आसार हैं जिनमें महामारी के दौरान वायरस के शिकार होने का ज़्यादा जोखिम है.

POLARITY: -0.5

STATUS: NEGATIVE

Fig -3: Snapshot of Implementation

5. CONCLUSION

We have used a set of natural language processing technique to classify a Hindi text or article as positive, negative or neutral based on polarity. As the rise in Hindi user-generated content is increasing in various fields the aim of our project is to classify Hindi news article efficiently to provide better clarity to the user. Sentiment analysis has helped in analyzing the attitude and sentiments of writer or the news that they express affirmatively, negatively or in any other way. Experimentation results specify that the proposed algorithm is performing well and achieved the accuracy of 83.29%.

REFERENCES

[1] A. Joshi, B. A. R, and P. Bhattacharyya, “A fallback strategy for sentiment analysis in Hindi: a case study”, International Conference Language Processing, 2010

[2] Nongmeikapam, Kishorjit, Sivaji Bandyopadhyay, DilipkumarKhangembam, Wangkheimayum Hemkumar, Shinghajit Khuraijam, "Verb Based Manipuri Aanalysis”,International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, pp. 113-119, June2014.

[3] Namita Mittal, Basant Agarwal, Garvit Chouhan, Prateek Pareek, and Nitin Bania (2013) “Discourse Based Sentiment Analysis for Hindi Reviews” P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 720–725, 2013.

[4] Balamurali A R,Aditya Joshi, Pushpak Bhattacharyya Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets.

[5] Richa Sharma, Shweta Nigam, Rekha Jain, (2014), ―Polarity detection of Movie Reviews in Hindi Language‖ in International Journal on Computational Sciences and Applications (IJCSA) Vol.4 No.4.

[6] Richa Sharma1, Shweta Nigam2 and Rekha Jain (2014b) “Opinion Mining in Hindi Language: A Survey” International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.2, March 2014.

[7] https://www.bbc.com/hindi/international-52580450

[8] www.aclweb.org/anthology/P14-2063

[9] https://data.mendeley.com/datasets/bsr3frvvjc/1