International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4

ISSN: 2454-4248
536 – 540

Hindi Sentiment Analysis

Sumedha Ubale, Ankita Sarang, Kajol Wadye, Prof. Nita Patil
[1,2,3,4]*Datta Meghe College Of Engineering,*
[1,2,3,4]*Airoli, Navi Mumbai.*

Abstract **–**With the evolution of web technology, a huge amount of data is present on the web. In addition to exploring the resources present on web, the users also provide feedback thus generating additional useful data. Thus mining of data and identifying user sentiments is the need of the hour. Sentiment analysis is the natural language processing task that mines information from various text forms such as blogs, reviews and classify them on the basis of polarity such as positive, negative or neutral. Hindi is the national language of India and is spoken by 366 million people across the world. The percentage of web content in Hindi is growing at lightning speed. A lot of research in opinion mining is carried out in English language but there are not many instances of research done in Hindi language. In this paper we have proposed a strategy for classifying given Hindi texts in to different classes and then extract sentiments in terms of positive, negative and neutral for identified classes. Naive Bayes, Modified Maximum entropy are used for classification and HindiSentiWordNet (HSWN) is used to determine the polarity of individual class.

Keywords: *Sentiment analysis, Machine Learning, HindiSentiWordNet (HSWN).*

*****

1. INTRODUCTION

With recent development of web 2.0 and Natural Language Processing, use of regional is grown for communication. As the internet is reaching more and more people within the world, there is tremendous growth in web content of other languages. Most of the research work is done in opinion mining of text in English language, very little work is done for other languages. Hindi is the fourth most spoken language in the world. The user generated Hindi content on webRSS is increasing rapidly. Therefore an efficient sentiment analysis for Hindi language is needed.

In our approach we are classifying a Hindi document into multiple classes and then extracting sentiments in the form of positive, negative and neutral from respective classes. Classification is the technique in which we identify which set of categories a new observation belongs to and this done on the basis of training set of data which contain observations whose category membership is already known. Multiclass classification is classifying observations into more than three classes.

Section I gives introduction of this topic. Section II gives brief description of sentiment Analysis in English as well as the work done in Indian Regional language like Hindi etc. Section III includes proposed architecture. Section IV describes methodology used in this paper. Section V describes the experimentation and evaluation in detail. At last Section VI gives conclusion and future scope of this paper.

2. RELATED WORK

There are many researches done in Sentiment Analysis by researchers in various languages such as English, Hindi, Bengali, Japanese, etc since 2001.

M. Farhadloo and E. Rolland [1] proposed Multi-Class Sentiment Analysis using the techniques of Clustering and Score Representation for English language. In this paper , they proposed improved methods for sentiment analysis at the aspect level by using a bag of nouns instead of bag of words to improve the clustering results and thus leading to more accurate sentiment identification.

A. Joshi, B. A. R, and P. Bhattacharyya [2] proposed a fall back strategy for sentiment analysis of Hindi language. In this model three 5approaches were used- In-language translation, machine translation resource based sentiment analysis for Sentiment analysis of Hindi language .In the first approach a sentimental annoted corpora has been developed in the Hindi movie review domain and it involves a training classifier to classify a new document in Hindi language. In the second approach, a classifier trained on standard English movie reviews has been used to translate the given document into English. In the third approach a lexical resource called HindiSentiWordNet is developed and majority score based strategy is implemented to classify a given document.

Kisorjit, Bandyopadhyay proposed a verb based approach for Sentiment analysis for Manipuri language. [3]. In this model an unsupervised learning approach called CRF (Conditional Random Field) is used. They also proposed the same model for Bengali language.

Aditya Joshi, Balamurli [4] proposed cross lingual sentiment analysis for Indian Languages.

K. M. Anil Kumar et al [5] proposed a model for retrieving user's sentiments from Kannada Web documents. Machine translation was used to translate the English reviews into Kannada, further POS tagger is used to implement adjective analysis and Turney algorithm which focuses on pair of POS. The polarity is considered as the difference between the positive and negative counts. If the value results more than zero then considered as positive, less than zero then negative else neutral.

 ISSN: 2454-4248
536 – 540

Keretna, Lim and Creighton [6] have worked on recognizing named entities from a medical dataset containing informal and unstructured text. For this, they combine the individual results of Conditional Random Field (CRF) classifiers and Maximum Entropy (ME) classifiers on the medical text; each classifier trained using a different set of features. CRF concentrates on the contextual features and ME concentrates on the linguistic features of each word. The combined results were better than the individual results of both the classifiers based on Recall rate performance measure.

Recent research by Fragos, Belsis and Skourlas [7] also concludes in favor of combining different approaches for text classification. The methods that authors have combined belong to same paradigm – probabilistic. Naïve Bayes and Maximum entropy classifiers are chosen to test on the applications where the individual performance is good. The merging operators are used above the individual results. Maximum and Harmonic mean operators have been used and the performance of combination is better than the individual classifiers.

3. PROPOSED SYSTEM

1. Build training data by using some Hindi text corpus
2. Identify the sentiments for text corpus.
3. Build classification model for predict the sentiment
4. Apply classification model on new test data.

3.1 Description of the system in short:

We will be following steps to implement the sentiment classification for Hindi Language

1. Read Hindi text file (UTF-8)
2. For each line, the sentiment score is identified as follows:
a. Identify each word from line and remove non-Hindi characters if any.
b. Each word is tokenized. c.
Stop words are removed.
d. Every word is reduced to its root word.
3. Build a classifier by correlating words and sentiment score:
a. Each word is classified into positive or negative class and also occurrence of word is calculated.
b. Then the probability of each word in that class is calculated.
4. Use Naive Bayes and Gini Index for classification.
5. Build the training data set.
6. Apply model on new test data and compare the results.

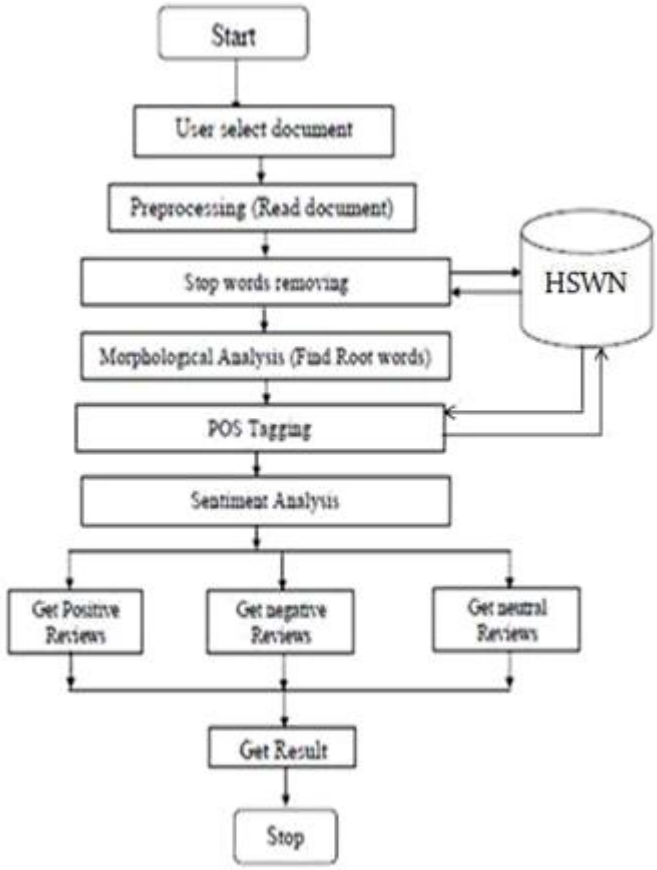


Fig. System Flowchart

4. MODULE DESCRIPTION

1. Validation:

We will validate whether the text is Hindi or Non-Hindi. For validation purpose we will be using HindiSentiWordNet. Along with that special characters like? / @ $ % will be removed.

2. Tokenization:

Sentence is divided into number of tokens with the help of java string tokenizer.

4. Stop Word Removal:

Stop words are the words which do not convey any meaning they are just used to complete the sentence. For example,

के,भि,वह,है

Here all the stop words will be removed.

5. Morphological analysis: Here
we will find root words.
For example,

 ISSN: 2454-4248
536 – 540

6. POS Tagging:

Part of speech will be assigned to each word.
Whether it is Noun, Pronoun, Verb or Adjective etc.
For example,

//NNP //JJ //NN

5. CLASSIFICATION

1. Naïve Bayes Classifier

The Naive Bayes classifier algorithm is most simple probabilistic models that generate data based on the assumption that "Given the context of the class, all attributes of the text are independent to each other." The Naïve Bayes technique begins the procedure by accepting text documents as word counts. In the next step, it calculates the class conditional probability which is then followed by the classification probability or posterior probability which is to be used by the trained classifier to predict the class of any document.

2. Modified Maximum Entropy Classifier

The ME Classifier is based on three concepts-Weights, features and Prediction Probability.
We can compute the weights using the Gini Index (weighting method) instead of using the conventional method in which we optimize the objective function in Maximum Entropy-Gini Index

3. Proposed Combined Classifier

Stage by stage representation of the classification process is illustrated in Fig.1. The proposed classification process consists of the following stages

Pre-processing stage
Feature Extraction stage
Individual Classification stage
Combining Classification stage
Final Results

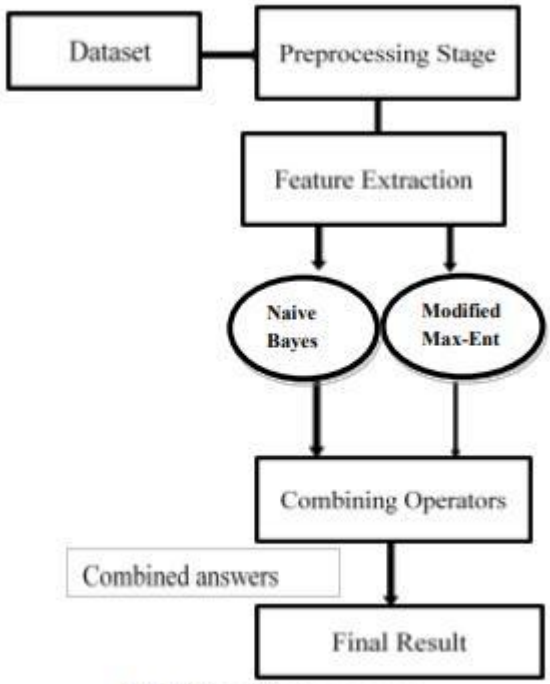


Fig 1: Classification process

The classification process starts with preprocessing of text to make it ready for the classification followed byextracting relevant features through Global Feature selection (GFS) method. It ranks the features according to their importance and the top K features are extracted. After the feature extraction process, Naïve Bayes (NB) and Maximum Entropy (ME) classifiers are used individually for classification. The later stage combines both the classifiers using three combination operators: Average, Harmonic Mean and Max. Combining operators are used for compensation of errors in each classifier and performance improvement.

$$Average(d) = \text{avg}(NB(d), ME\ (d))$$

$$Max(d) = \max(NB(d), ME\ (d)$$

$$Harmonic(d) = \frac{(2.0 * NB(d) * ME(d))}{(NB(d) + ME(d))}$$

The results of the combination then give the final result.

4. Algorithm for Classification

INPUT: Training data which includes term frequencies, class labels, test document d, number of classes C
OUTPUT: For a document d Prediction class C
Step 1: Using Training Data and class labels train the Naïve Bayes Classifier's class conditional probability .

 ISSN: 2454-4248
536 – 540

Step 2: For every class Compute posterior class probability of Naïve Bayes Classifier

Step 3: Train the maximum entropy classifier.ME Classifier: this can be done by using weighing schemes such as CHI-Square, DIA factor, Gini Index or CMFS and feature function as well.

Step 4: For every class using Maximum Entropy compute posterior probability.

Step 5: For every class compute probability d

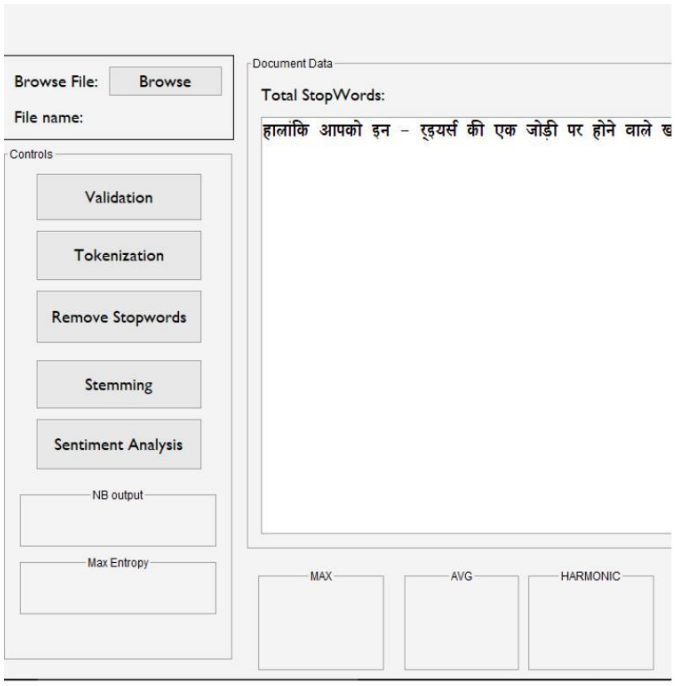
(|) = COMB((|) ,(|))

Where
COMB denotes any one of the operator Average. Max or Harmonic Mean .

Step 6:prediction of class d which is maximum.

6. IMPLEMENTATION RESULTS

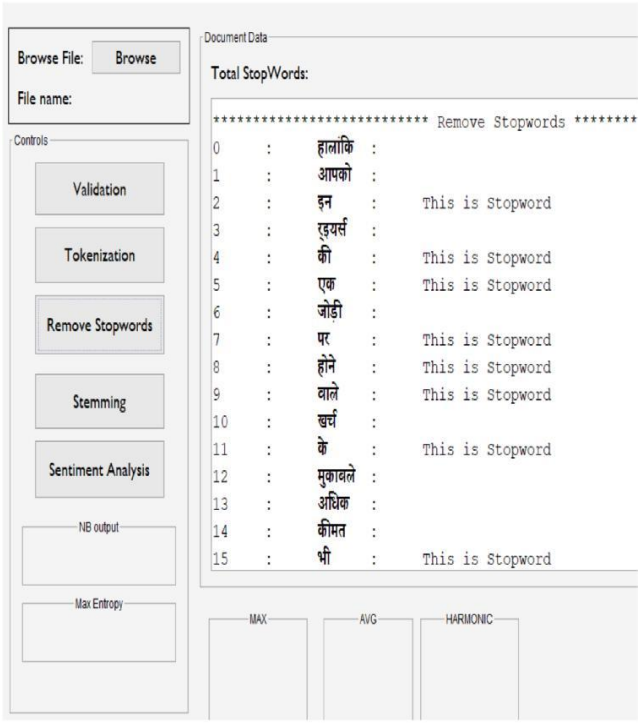
Step 1: Insert the text file or write in the text area.



Step 2: Validation

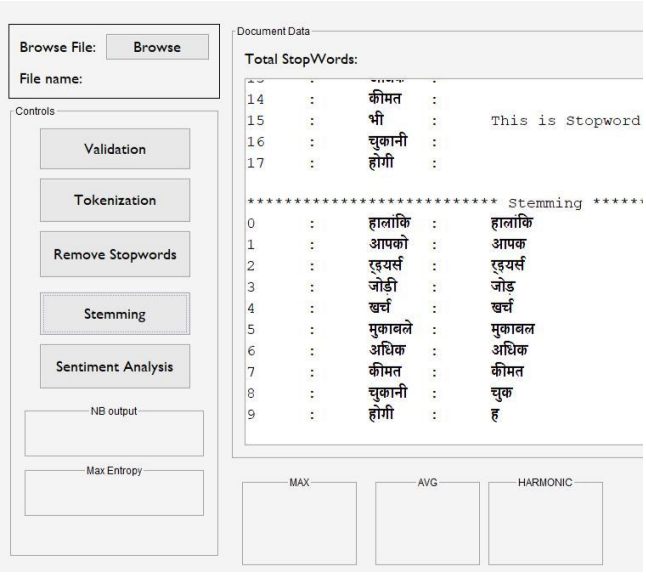
Step 3: Tokenization

Step 4: Remove Stopwords.

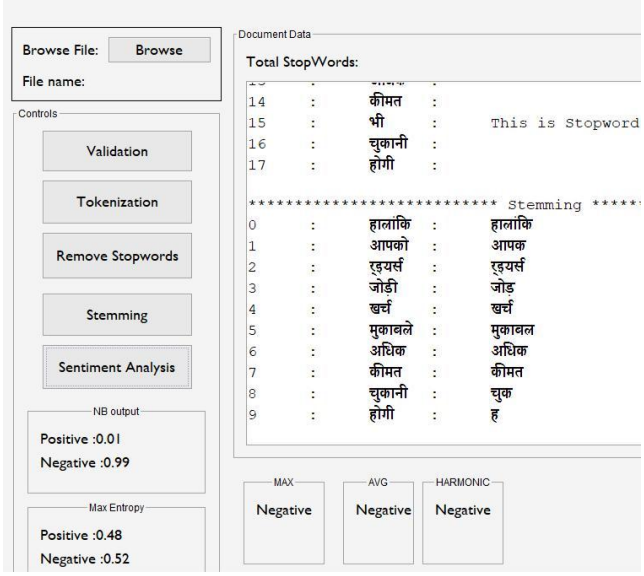


 ISSN: 2454-4248
536 – 540

Step 5: Stemming



Step 6:



6. CONCLUSION

In this project, we will be using a set of machine learning techniques to classify a hindi text as positive, negative or neutral based on polarity.The naïve bayes technique along with maximum entropy and gini index gives good enough results and better performance as well . The training dataset can be increased to improve the feature vector related sentence identification process. We are using HindiSentiWordNet for the summarization of the document. It may give better visualization of the content in better manner that will be helpful for the users. As the rise in Hindi user generated content is increasing in various fields the aim of our project is to classify Hindi texts efficiently to provide better performance to the user.

8. REFERENCES

[1] Mohsen Farhadloo, Erik Rolland," Multi-Class Sentiment Analysis with Clustering and Score Representation", IEEE 13 th International Conference on Data Mining Workshops, pp. 904-912, December 2013

[2] A. Joshi, B. A. R, and P. Bhattacharyya, "A fall-back strategy forsentiment analysis in Hindi: a case study", International Conference Language Processing, 2010

[3] 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.

[4] Balamurali A R, Aditya Joshi, Pushpak Bhattacharyya," Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets", COLING 2012, pp. 73-8, December 2012.

[5] K. M. Anil Kumar, N. Rajasimha, M Reddy, A. Rajanarayana, K. Nadgir, "Analysis of Users' Sentiments from Kannada Web Documents", Eleventh International Conference Communication Networks, vol. 54, pp. 247-256, August 2015