SENTIMENT ANALYSIS OF INDIAN LANGUAGE

Sheetal Sharma[1], S K Bharti[2], Raj Kumar Goel[3]

[1]M. Tech Student, Department of Computer Science and Engineering, Noida Institute of Engineering & Technology, Greater Noida, U.P., India

[2,3] Professor, Department of Computer Science and Engineering, Noida Institute of Engineering & Technology, Greater Noida, U.P., India

---***---

Abstract - This project contains the sentiment analysis on Hindi language with the help of data mining and machine learning techniques. The sentiment analysis was done on Hindi news. In addition to resource creation, we take up the task of sentiment classification in Hindi Language. It aims to determine the attitude of a sentences with respect to some topic or simply the contextual polarity of a document. Early work in this area was done by Turney and Pang, who applied different methods for detecting the polarity of product and movie reviews. This project contains the sentiment analysis on Hindi language with the help of data mining and machine learning techniques. The sentiment analysis was done on Hindi news. The results show the polarity of the sentiments in Hindi news sentences as positive and negative. We aim at analysing the merits and demerits of each of the above approaches across the different genres for the sentiment classification task. We discuss in detail the problems and the issues while working with the user-generated content (reviews and news) in Hindi language.

Key Words: Sentiments, Polarity, Hindi Language, Analysis

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 on the Web.

Client created content is a vital wellspring of data to mine the feeling/conclusion of individuals on various items and administrations. The creating innovation easily of reachability and better availability has prompt far reaching utilization of websites, discussions, e-news, surveys channels and the person to person communication stages, for example, Facebook, Twitter. These long range informal communication stages has exponentially expanded the measure of data produced on everyday schedule. In this way mining the information and recognizing client conclusions, wishes, different preferences is one of a critical assignment that has pulled in the focal point of research group from a decade ago. The internet assumes a urgent part in get-together general feeling, these suppositions assume a critical part in settling on business related choices.

To obtain the factual and subjective information on companies and products, analysts are turning towards web to gather information. Extracting public opinion from this information is a major task. Industrialists spend a large chunk of their revenue on business intelligence to read minds of general public and interpret what they think about their product(s). Sentiment analysis tries to mine information from various text forms such as reviews, news, blogs and classify them on the basis of their polarity as positive, negative or neutral.

Naive Bayes is a very simple probabilistic model that tends to work well on text classifications and usually takes orders of magnitude less time to train when compared to models like support vector machines. A high degree of accuracy can be obtained using Naive Bayes model, which is comparable to the current state of the art models in sentiment classification.

- Sentiment Analysis, also known as Opinion Mining
- The process of determining the emotional tone behind a piece of text, used to gain understanding of attitudes, opinions and emotions.

Sentiment analysis is a complicated problem but experiments have been done using Naive Bayes, maximum entropy classifiers and support vector machines. Pang et al. found the SVM to be the most accurate classifier in. We present a supervised sentiment classification model based on the Naive Bayes algorithm.

Goals

1. To do Sentiment Analysis on Indian Language.

2. To plot graph between features v/s accuracies and find the best accuracy.

Motivation and Applications

- Insight from social data, application in business intelligence like marketing

- Cross domain application like sociology, psychology or administration

- Useful in feedback and recommendation systems

2. LITERATURE REVIEW

Namita Mittal, Basant Agarwal, Garvit Chouhan, Prateek Pareek, and Nitin Bania (2013) studied on the hindi 1anguage content which was also growing very fast online. Sentiment classification research has been done mostly for English language. In this paper, it is investigated that how by proper handling of negation and discourse relation may improve the performance of Hindi review sentiment analysis. Experimental results show the effectiveness of the proposed approach.

Komal Garg and Preetpal Kaur Buttar (2015) depicted Sentiment investigation (SA) is one of the quickest developing examination zones in natural language processing, making it trying to monitor every one of the exercises in the zone. Increment in user-generated content (UGC) has given an imperative angle to the scientists, businesses and government(s) to mine this data. It examined general feeling score of entered hindi content t0 arrange each sentence as positive, negative and impartial. In this postulation, we take a shot at the slant examination by decaying a calculation that distinguishing the supposition as indicated by proposed rules in view of places of conjunction, refutation and perspectives (things).

Shanta Phani, Shibamouli Lahiri and Arindam Biswas (2016) in this paper, depicted the consequences of conclusion investigation on tweets in three Indian languages – Bengali, Hindi, and Tamil. We utilized the as of late discharged SAIL dataset and got cutting edge brings about every one of the three dialects. Points by point highlight were examined and blunder investigation had been accounted for, alongside expectations to learn and adapt for Hindi and Bengali.

3. Data Collection for Indian language

To perform sentiment analysis in Indian language, the data set has to be prepared first. To prepare the data set, large numbers of Hindi news sentences were collected from the Web. There are lots of websites like which contain Hindi content. Here, Hindi news sentences were collected from the Hindi newspapers website. But before applying as an input, the collected data first preprocessed. After preprocessing the reviews were applied as an input. All the data after getting preprocessed, it is taken as input for the analysis system and using algorithms, we get the results in the form of output as shown in Figure 1.

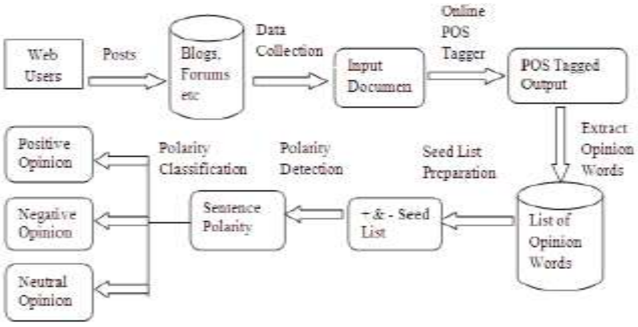

Fig -1: Indian Languages Sentiment Analysis System

3.1 Part of Speech Tagging

POS tagging is very important for sentiment and sentences mining. POS tagging is used to determine the opinion words and features in the reviews. POS tagging can be done manually or with the help of POS tagger. POS tagger tag all the words of reviews to their appropriate part of speech tag. Manual POS tagging of the reviews takes lots of time. Here, Online POS tagger of Hindi is used to tag all the words of reviews.

3.2 Polarity detection of reviews

In the last phase, the polarity of the collected reviews is determined with the help of seed list and Hindi dictionary. The polarity of the reviews is determined on the basis of majority of opinion words, if positive words are more in the review than the polarity of the review is positive otherwise it is negative. If positive and negative words are equal in a review the polarity is neutral. As negation is also handled in this approach, so if the opinion word is followed by not then the polarity of review is reversed. e.g. the sentence.
अभिजीत यह गाना अच्छा नहीं गा पाये | Here, the opinion word is ‘’ ‘अच्छा’ which is followed by ‘ ‘नहीं’ shows negative polarity.

4. EXPERIMENTS & RESULTS

Experiment is conducted on movie reviews. Movie reviews were collected from several websites contain Hindi reviews. Reviews were applied as input to the system which classifies these reviews and determine the polarity of these reviews and present the summarized positive and negative results which prove to be helpful for the users. Input reviews were also classified by us to determine how well the system classified the reviews as compared to human judgement. Three evaluation measures are used on the basis of which system performance is computed, these are:

- Precision
- Recall
- Accuracy

The common way for computing these measures is based on the Polarity matrix shown in

Table 1: Polarity Matrix

Sentences	Predicted positives	Predicted negatives
Positive Sentences	# of True positive sentences	# of false negative sentences
Negative Sentences	# of false positive sentences	# of True Negative sentences

In the Hindi sentiments orientation systems is performed on the Hindi news sentences domain. The experiments have been performed by using of Hindi news sentences domain. Table 2 presents the data of Hindi sentiments in terms of sentences, train positive and negative as input data.

Table 2: Input Data

Measure	Results
Positive Sentences	1482
Negative Sentences	1503
Train Positive	1185
Train Negative	1202
Test positive	297
Test Negative	301

Table3: Output data

Measure	Results
Total no of times positive word occurred in data set	13731
Total no of times negative word occurred in data set	15546
Total no of features	2615
optimal value of k	850
Best accuracy	0.664041994751

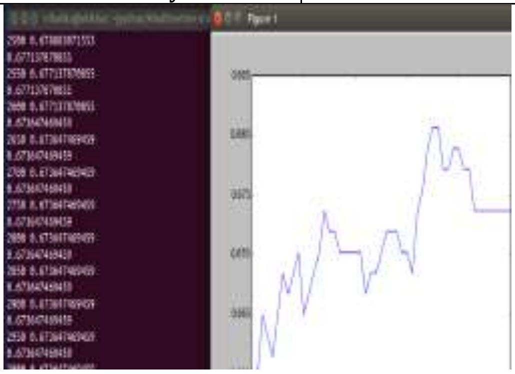

Figure 3.1: Graphical representation of results in the form of polarity of words

5. CONCLUSIONS

Sentiment Analysis using Data Mining is an emerging research field and is very important because human beings are largely dependent on the web nowadays. The rise in user-generated content in Hindi language across various genres- news, culture, arts, sports etc. has open the data to be explored and mined effectively, to provide better services and facilities to the consumers. The best k was found at instance 850 and accuracy was found about 0.664041994751 at the given dataset (3000 Hindi News Sentences)

REFERENCES

[1]. Deepali Mishra, Manju Venugopalan and Deepa Gupta (2016) “Context Specific Lexicon for Hindi Reviews” Procedia Computer Science, 93, 554 – 563.

[2]. Richa Sharma,Shweta Nigam and Rekha Jain (2014) “Polarity Detection Of Movie Reviews in Hindi Language” International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.4, August 2014.

[3]. Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya (2014) “Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi”.

[4]. Namita Mittal and Basant Agarwal (2013) “Sentiment Analysis of Hindi Review based on Negation and Discourse Relation” International Joint Conference on Natural Language Processing, pages 45–50, Nagoya, Japan, 14-18 October.

[5]. Shanta Phani, Shibamouli Lahiri and Arindam Biswas (2016) “Sentiment Analysis of Tweets in Three Indian Languages” Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, pages 93–102, Osaka, Japan, December.

[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]. Komal Garg and Preetpal Kaur Buttar (2015) “Aspect based Sentiment Analysis of Hindi Text Review” International Journal of Advanced Research in Computer Science, Volume 8, No. 7, July – August 2017

[8]. 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.