Sentiment Analysis in Marathi Language

[1]Snehal V. Pawar, [2]Prof. Swati Mali
Dept. Of Computer Technology
K. J. Somaiya College Of Engineering
VIDYAVIHAR, MUMBAI
[1]*snehal.vp@somaiya.edu*, [2]*swatimali@somiaya.edu*

Abstract— Sentiment analysis is inevitable in current era. Internet is growing day-by-day. Now-a-days everything is online. We can shop, buy, and sell online. People can give feedbacks / opinions on the internet. Customers can compare among various products by analyzing the product reviews. As more and more people from different age groups and languages are becoming new internet users, we need it in regional languages. Till date most of the work related to sentiment analysis has been done in English language. But when it comes to Indian languages, not much research has done except for few languages. This paper mainly focuses on performing sentiment analysis in one of the Indian languages i.e. Marathi.

Keywords- *sentiment analysis, SVM, NB,Max.Entropy*

*****

I. INTRODUCTION

Sentiment analysis is an ongoing research field. In Sentiment analysis based on the sentiment value it is decided whether the sentence is positive, negative or neutral. This helps a lot when you need to rely on people's opinion. For example, if a mobile company launches a new mobile phone, it needs to know whether the customers like the product or not. They need to know that their product has fulfilled the customer's requirements or it needs more improvements. The easier way to understand that is to focus on the reviews / feedbacks. But reading all the feedbacks /reviews is itself a difficult task and concluding something from them adds up to the pile. If there is some technique or algorithm which analyses all the reviews and tell you whether a review is positive or negative, it will save a lot of time and overhead. Also if the algorithm tells you that how much positive or negative reviews you received for a particular product or for which aspect it got positive reviews and which aspects need improvements then it will become easier for the company or manufacturer to understand the customer's need and that's where Sentiment analysis come into picture.

Sentiment analysis techniques are broadly categorized into two approaches; machine learning and lexicon based approach. In machine learning approach, machine learning algorithms are used while in lexicon approach depends on the lexicon which consists of pre-defined sentiment words. Lexicon based approach is further divided into corpus-based and dictionary based approach.

If there is an algorithm which extract all the reviews related to a product and analyze them and tell you whether the product is good or bad or the algorithm will analyze the reviews of a movie and can tell you whether it is a hit or flop, then it will reduce a lot of overhead.

That is where sentiment analysis comes into existence. It uses various techniques to analyze the given data and extract sentiments out of it. The first step in sentiment analysis is to gather the data. The second step is to clean and pre-process the data. Then the data is given to the sentiment analysis techniques for further processing. At the end of the process, it assigns polarity to the data based on which it is determined that the data is either positive or negative.

As internet is growing day by day and people are expressing their opinions in various languages, we need to find an approach to extract sentiments out of them. Sentiment analysis is very important to understand the people opinions, but English isn't everyone's forte. Some people do want to write/express opinion in their mother tongue. To perform Sentiment analysis on this data, we do not have much resource for Marathi language available, as most of the work of Sentiment analysis is done in English language. Therefore this is a basic approach to perform sentiment analysis on data in Marathi language.

To perform sentiment analysis in Marathi language, we are using lexicon based techniques which requires lexicon containing positive words and negative words along with their polarity. Later they will be used to analyze the sentiment of the sentence. There will be a training set to train the classifier and "test data" to evaluate the performance.

The paper consists of the process of Sentiment analysis. Section II describes the previous work done by other authors along with their techniques and results. The motive behind the implementation of this approach is given in Section III. Section IV explains the proposed system, techniques used

in proposed system and issues related to sentiment analysis. The architecture is explained in section V with results of proposed system.

II. PREVIOUS WORK

Business holders want to know the requirements of the customers, also the opinions about the product. Customer feedback is there from a long time. Today in technology era, this feedback is emerged as sentiment analysis. Sentiment analysis started in late 2000s, but it is been in effect since 2004 in product reviews area. There are a lot of people who have done research on sentiment analysis, its process and its various techniques. There are some authors who have worked on the languages other than English. The ones who have performed Sentiment analysis in Indian languages, few of them are listed below.

The [1] is about movie review data in Hindi. It performs opinion mining at document level. It classifies documents in three categories viz., positive, negative and neutral. It has used Machine learning approach and Pos tagging. In the POS tagging only adjectives are concerned. Authors have performed both the methods in Machine Learning i.e. supervised and unsupervised. They have taken care of the negation as well. They have achieved 87.1% accuracy.

In [2], authors have done sentiment analysis on mixed language sentences. Here they have considered only two languages i.e. Hindi and English. The analysis has been done in phrase as well as sub-phrase level of the sentence. Grammatical transitions are taken into consideration while predicting the overall sentiment of the sentence. They have able to achieve accuracy up to 91%.

[3]This is mainly a survey paper. It gives various methods used in opinion mining. It is the summary of what work has been done related to opinion mining in Hindi language. It describes different challenges that need to be overcome while performing opinion mining in Hindi language. Authors have accepted that it is not easy to perform sentiment analysis in Hindi as it is a natural language and lack of resources.

There is a case study written on sentiment analysis using mobile phone reviews in Kannada language [4]. To gather the product reviews lexicon based approach has used. To determine the polarity naive Bayes classifier has applied. There is very less amount of work done in Kannada sentiment analysis. In this paper, they have used aspect based sentiment analysis. They have managed to get approx. 65% accuracy with their proposed model.

[5]In this paper sentiment analysis is done on movie reviews in Malayalam language. They have used rule based approach for sentiment analysis. Negation handling has done in order to extract sentiment from the review. The corpus is made from Malayalam websites. They have got 85% accuracy.

In the second paper [6] by the same authors as [5], they have taken the research work further ahead. They have used machine learning with rule based approach which gave them 91% accuracy by keeping the same corpus. They have used two techniques Support vector machine (SVM) and Conditional random fields (CRF). Authors have concluded that SVM is better than CRF.

In [7], authors have analyzed Bengali language for sentiment analysis. They have used SentiWordNet and WordNet to build the corpus. First it finds part of speech and then on the basis of the adjective they assign polarity to it. They have come up with their own method for sentiment extraction. They have achieved up to 90% accuracy in their results.

Another paper [8] has performed sentiment analysis in Bengali language using the horoscope in the newspaper. Approx. 6000 sentences have analyzed. They have used various machine learning techniques along with unigram and bigram methods. They have experimented SVM with unigram features without removing stop words which gave them 98.7% accuracy.

III. MOTIVATION

Marathi is an Indian language. It is the official language of Maharashtra. There were 73 million speakers in 2007; Marathi ranks 19th in the list of most spoken languages in the world. Marathi has the fourth largest number of native speakers in India, after Hindi, Bengali and Telagu in that order. Marathi employs agglutinative and analytical forms. Marathi uses many morphological processes to join words together, forming compounds. Now-a-days various Marathi typing software are widely used and display interface packages are now available on Windows and Linux. Many Marathi websites, including Marathi newspapers, have become popular. People sometimes use regional language to express their opinions on internet, Marathi is one of them. To perform Sentiment analysis on this data, we do not have much resource available, since not much work has been done on this topic. Most of the work of Sentiment analysis is done in English language. Therefore this is a basic approach to perform sentiment analysis on data in Marathi language.

IV. PROPOSED SYSTEM

There are various techniques for Sentiment analysis, broadly categorized into two approaches; machine learning and lexicon based approach. In machine learning approach machine learning algorithms are used while lexicon approach depends on the lexicon which consists of pre-defined sentiment words.

In this project, lexicon based approach is used which requires a lexicon consisting positive and negative words. A dataset is created comprise of positive and negative words consisting 100 positive and 100 negative words. Positive words are having polarity 1 and negative words are having polarity 0.

A. *Support Vector Machine*

Support Vector Machine is a supervised machine learning algorithm for classification or regression problems where the dataset teaches SVM about the classes so that SVM can classify any new data. It works by classifying the data into different classes by finding a line which separates the training data set into classes. In this algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. SVM offers best classification performance on the training data. SVM renders more efficiency for correct classification of the future data. The best thing about SVM is that it does not make any strong assumptions on data. It does not over-fit the data. It is effective in high dimensional spaces. It is memory efficient. Though, it doesn't perform well, when we have large data set because the required training time is higher. There are several applications such as Classification of images, Hand-written characters recognition, The SVM algorithm has been widely applied in the biological and other sciences. SVM is commonly used for stock market forecasting by various financial institutions. For instance, it can be used to compare the relative performance of the stocks when compared to performance of other stocks in the same sector. The relative comparison of stocks helps manage investment making decisions based on the classifications made by the SVM learning algorithm.

B. *Naïve Bayes*

A classifier is a function that allocates a population's element value from one of the available categories. For instance, Spam Filtering is a popular application of Naïve Bayes algorithm. Spam filter here, is a classifier that assigns a label "Spam" or "Not Spam" to all the emails. Naïve Bayes Classifier is amongst the most popular learning method grouped by similarities that works on the popular Bayes Theorem of Probability- to build machine learning models particularly for disease prediction and document classification. It is a classification technique based on Bayes theorem. Naive Bayes classifier assumes that a particular feature is independent. Naive Bayes model is easy to build and particularly useful for very large data sets. Naive Bayes is known to outperform many classification methods.

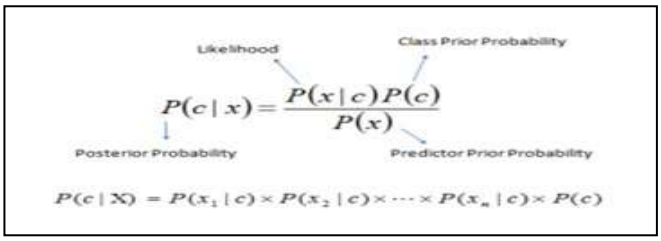

Figure 1. Naïve Bayes probabalistic model.

Naïve bayes is easy and fast as it needs less training data. It can be used for making predictions in real time. This algorithm is also well known for multi class prediction feature. It can be used for Text classification, Spam Filtering, Sentiment Analysis. Limitation of Naive Bayes is the assumption of independent feature. In real life, it is almost impossible to get independent features.

Applications include Spam filtering, Classify documents based on topics, Sentiment analysis, Information retrieval, Image classifications, Medical field, Document Categorization. Naïve Bayes Algorithm is also used for classifying news articles about Technology, Entertainment, Sports, Politics, etc.

C. *Maximum Entropy*

The Max Entropy classifier is a probabilistic classifier. Max Entropy does not assume that the features are independent of each other. The classifier is based on the principle of Maximum Entropy from all the models that fit the training data, selects the one which has the largest entropy. The principle of maximum entropy states that, subject to precisely stated prior data, the probability distribution which best represents the current state of knowledge is the one with largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more. Due to the minimum assumptions that the Maximum Entropy classifier makes, we regularly use it when it is unsafe to make any such assumptions. The Max Entropy requires more time to train comparing to Naive Bayes**.**

D. *Issues related to Sentiment Analysis*

There are some issues related to sentiment analysis as follows:

1. **Named entity recognition**: - Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The problem is we do not know what a person actually wants to say. For example "Bank of America" is one name which has "America" as substring which may be considered as noun while tagging.
2. **Sarcasm**: - In sentiment analysis sarcasm plays an important role. It may have positive words or negative words but the meaning of the sentence is not as it seems to be. Detection of sarcasm is important in sentiment analysis to get the exact meaning of the sentence but it is a serious technical challenge.
3. **Anaphora resolution**: - The problem is to understand what a noun or pronoun refers to such as "we watched the movie and went to dinner, it was awful". What does "awful" refer to?

4. **Polarity**: - In general the polarity of particular word or sentence is positive or negative or neutral. But how much positive or negative is another question. "Good" and "Best" both are positive but second one is a stronger sentiment than the first one.
5. **Use of abbreviations poor spelling punctuation grammar**: - Due to these issues it is difficult to identify the sentiment or aspect on which opinion is expressed.
6. **Sentiment lexicon acquisition**: - There are a lot of words available in the dictionary for English language. But opinions can be expressed in the form of words, emoticons, slang words etc. having all these words in the lexicon is not that easy. Also multi-lingual words can be there. Therefore finding sentiment in such opinions is a difficult task.
7. **Negation handling**: - Only writing not in the sentence does not make it a negative sentence such as "The food is not bad" has positive opinion but it may consider as negative because of the "not" word.
8. **Aspect based Sentiment Analysis**: - We should know on which aspect the opinion is expressed so as to evaluate the sentiment. Rather than evaluating sentiment of whole sentence finding the aspect is more important.

There are other overheads also such as:

1. On internet there is a lot of spam and fake messages. We need to eliminate the fake reviews to get efficient sentiment analysis.
2. There are sentiment analysis tools available but they are expensive which cannot affordable by common person.
3. The sentiments are domain dependent. Therefore features which give good performance in one domain may fail in other domain.

V. ARCHITECTURE

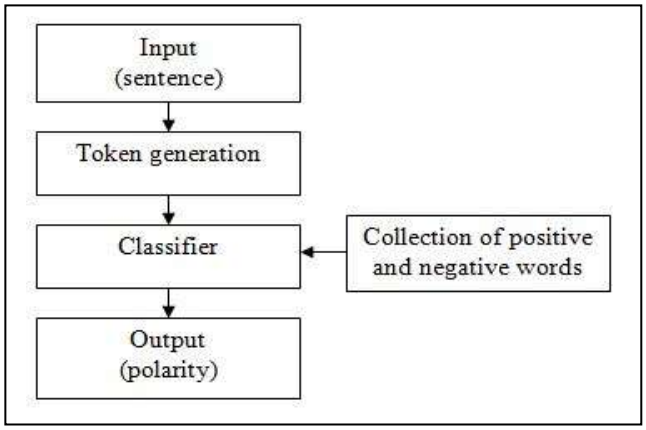

Figure 2. Architecture of proposed system.

Steps involved in proposed system-

Step 1: A sentence is given as an input.
Step 2: Each word in the given sentence is extracted
Step 3: The extracted word is searched in the dataset
Step 4: If found its polarity is collected
Step 5: If polarity is 1 then sentence is positive
Step 6: If polarity is 0 then sentence is negative

As given in the architecture diagram, a sentence is analyzed based on the sentiment words it contains regardless of other words. For this purpose, only adjectives are considered. If the sentiment word is found in our positive dataset, then the sentence is considered as positive. If the sentiment word is found in our negative dataset, then the sentence is considered as negative. If not found in any then it is considered as neutral.

.

A. Results

A user interactive webpage is created which will ask user to enter a sentence. Results are shown in three tables' viz., positive words, negative words, neutral words. Those words matched with positive words of the dataset are considered as positive and those words matched with negative words of the dataset are considered as negative. Words not matching with either of positive words and negative words are considered as neutral.

TABLE I. EXAMPLES

Sentences	Polarities		
	Positive	***Negative***	***Neutral***
आमचे शेजारी भांडखोर आहेत		भांडखोर	आमचे, शेजारी ,आहेत
आकाश हा कपटी माणूस आहे		कपटी	आकाश, हा, माणूस, आहे
अर्जुन धनुर्विद्येत कुशल होता	कुशल		अर्जुन, धनुर्विद्येत, होता
आकाश प्रामाणिक मुलगा आहे	प्रामाणिक		आकाश, मुलगा, आहे
काश्मिर हे एक रमणीय ठिकाण आहे			काश्मिर, हे, एक, रमणीय, ठिकाण, आहे
ही कादंबरी नीरस आहे			ही, कादंबरी, नीरस, आहे

As per above examples, the words 'भांडखोर', 'कपटी' are having polarity '0' which indicates that they are present in 'negative_words' list, therefore considered as negative words and the sentences comprising such words are referred as negative.

Similarly 'कुशल', 'प्रामाणिक' both are having polarity '1' which indicates that they are present in 'positive_words' list,

therefore considered as positive words and the sentences comprising such words are referred as positive.

The word 'रमणीय' in 'काश्मिर हे एक रमणीय ठिकाण आहे' is positive, but it is not present in our 'positive_words' list therefore it is considered as neutral.

Similarly, the word 'नीरस' in 'ही कादंबरी नीरस आहे' is negative, but it is not present in our 'negative_words' list therefore it is considered as neutral.

VI. CONCLUSION AND FUTURE WORK

Different research papers are studied to understand how different techniques work and how they affect sentiment analysis under different circumstances. Based on the research done during the proposed work following conclusions are made -

- The project requires analyzing data in Marathi language.
- It is slightly difficult to process data in Marathi as it is a natural language and not much work has been done related to this topic.
- According to the results, the words which are not in the database are considered as neutral though they are sentiment words. Therefore the database should be as rich as possible for efficient results.

This concludes that not only the techniques but the resources are also more important for better results.
The dataset will be enhanced as part of future work. Also other algorithms and techniques will be implemented to increase the efficiency.

REFERENCES

[1] Jha, Vandana, N. Manjunath, P. Deepa Shenoy, K. R. Venugopal, and Lalit M. Patnaik. "Homs: Hindi opinion mining system." In *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*, pp. 366-371. IEEE, 2015.

[2] Sitaram, Dinkar, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. "Sentiment analysis of mixed language employing Hindi-English code switching." In *Machine Learning and Cybernetics (ICMLC), 2015 International Conference on*, vol. 1, pp. 271-276. IEEE, 2015.

[3] Sharma, Richa, Shweta Nigam, and Rekha Jain. "Opinion mining in Hindi language: a survey." *arXiv preprint arXiv:1404.4935* (2014).

[4] Hegde, Yashaswini, and S. K. Padma. "Sentiment Analysis for Kannada using mobile product reviews: A case study." In *Advance Computing Conference (IACC), 2015 IEEE International*, pp. 822-827. IEEE, 2015.

[5] Nair, Deepu S., Jisha P. Jayan, and Elizabeth Sherly. "SentiMa-sentiment extraction for Malayalam." In *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, pp. 1719-1723. IEEE, 2014.

[6] Nair, Deepu S., Jisha P. Jayan, R. R. Rajeev, and Elizabeth Sherly. "Sentiment Analysis of Malayalam film review using machine learning techniques." In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pp. 2381-2384. IEEE, 2015.

[7] Hasan, KM Azharul, and Mosiur Rahman. "Sentiment detection from Bangla text using contextual valency analysis." In *Computer and Information Technology (ICCIT), 2014 17th International Conference on*, pp. 292-295. IEEE, 2014.

[8] Ghosal, Tirthankar, Sajal K. Das, and Saprativa Bhattacharjee. "Sentiment analysis on (Bengali horoscope) corpus." In *India Conference (INDICON), 2015 Annual IEEE*, pp. 1-6. IEEE, 2015.