Prediction of Sentimental Analysis using Machine Learning Technique

Mr. Vinayak Hegde, Asst.Professor
Mr. Narendra Kamath G
Department of Computer Science
Amrita Vishwa Vidyapeetham, Mysuru campus

***Abstract* – It is very important to know whether the given text is positive, negative or an objective sentence. As a human it is very easy to classify the sentences by just understanding the intonation and usage of the text. But it is the challenging task for a computer program to classify the text. Even in some situations it is difficult for a human to categorize the text. The possible methodology which can be followed is the sentiment classification. We observed that a combination of methods like Naive Bayes classifier, effective negation handling, emphasizing words handling and feature selection results in a significant improvement in the accuracy. The proposed system can be an effective solution for the text categorization problems.**

***Keywords:-* Sentiment Classification, Naive Bayes classifier, Negation Handling, Emphasizing words handling, feature selection**

I. INTRODUCTION

Opinions are very important in each and every field of work, but it is important to classify the opinions in to the negative, positive or objective type of opinion. These opinions play a major role in market for the product review analysis and also in understanding whether a given text is in positive or negative context for the sentimental analysis of the text which is available on internet [8]. The opinions also plays a major role in politics where the politician wish to know how people actually feel about the present government, such that by analyzing the opinions of people, such that the future election results can be predicted accurately. It is very easy for a human to analyze a text and to predict the sentiment of the given text, by understanding how the words are used in the text or by understanding the context of the given text. In order to classify the opinions we can use a methodology called sentiment classification. For example, the sentence "It is a very good chance to kill him", is definitely a negative sentence. But the usage of the positive and emphasizing words good and very respectively will give a positivity in the sentence, but using the word 'kill', entire context of the sentence itself changes. So it is noticed that, only by considering the number of positive and negative words, we cannot draw a conclusion that the text is either positive or negative. It is necessary procedure to notice the words which are used before or after the used positive or negative words. Stating another example, "Unemployment increases the crime rate in the city", the expected result of this statement is definitely negative. In these types of text we have to identify the emphasizing words ("increases") and then we have to see to which type of context the emphasizing words are used. In the current context the emphasizing word 'increases' refers to the phrase crime rate, which means that the negativity is emphasized, which increases the negativity of the given text.

It is a complicated problem to classify the sentiments given the text, but experiments have been done using Naive Bayes classifier, maximum entropy classifiers and support vector machines. In this paper we present an efficient sentiment classification model with the combination of methodologies like Naive Bayes classifier, effective negation handling, emphasizing words handling and feature selection which results in a significant improvement in the accuracy of sentiment classification.

II. DEFINITION

Sentiment classification is a methodology which is used to categorize the text into positive, negative or objective sentiment [1] [16]. It can also be stated as a technique used to analyze the subjective information in the text and to categorize the text. A typical approach for sentiment classification is to use machine learning algorithms.

A. Machine Learning

Machine learning refers to the construction and study of the algorithms that can learn from the given dataset. These types of algorithms use the example inputs and use the same to make the predictions or decisions, rather than following strict static program instructions.

B. Supervised Learning

Supervised learning refers to generation of a function which maps the input to desired output, with given set of training data (Figure 1). Since it is a text classification problem, any supervised learning method can be applied, e.g., Naive Bayes classification, and support vector machines.

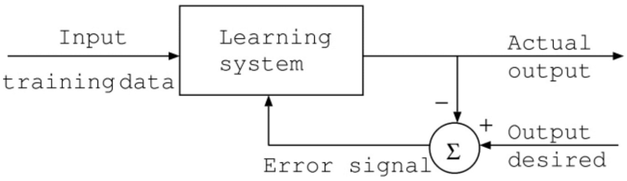


Figure 1 : Supervised Learning

C. Unsupervised Learning

Unsupervised learning refers to modeling a set of inputs, like clustering, labels are not known during training. Classification is performed using some fixed syntactic patterns which are used to express opinions (Figure 2).

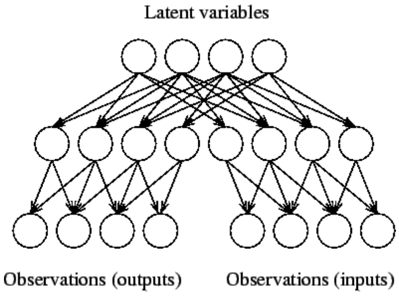


Figure 2: Unsupervised Learning

D. Semi-supervised Learning

Semi-supervised learning generate an appropriate function or classifier in which both labeled and unlabeled examples are combined (Figure 3).

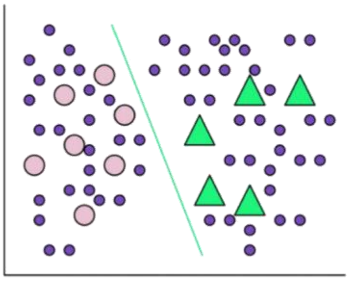


Figure 3: Semi-supervised Learning

III. LEVELS OF SENTIMENTAL ANALYSIS

The sentiment analysis can be performed at one of the three levels: the document level, sentence level, feature level.

A. Document Level

A document level of sentiment analysis, the whole document is considered as a single entity and the analysis is done on the whole document. Document is usually a combination of all types of sentences. In this approach towards sentiment classification, the main challenge is to extract informative text for inferring sentiment of the whole document. There may be conflicting sentiments if the document is complicated. The results generated in this level may not be accurate always. But the solution for the same can be the next level of sentiment analysis i.e. sentence level sentiment analysis.

B. Sentence Level

Sentences are the group of words arranged in a meaningful manner. This level of sentiment analysis can also be stated as a fine-grained level of document level of sentiment analysis. In this level we can check the polarity factor of the sentence and it is possible to categorize the sentence into one of the three possible categories of the sentiment i.e. positive, negative or objective. The challenge faced by sentence level sentiment classification is the identification features indicating whether sentences are on-topic which is kind of co-reference problem.

C. Feature Level

The feature level of sentiment analysis is considered to be the fine grained analysis model among all other models. The product features are defined as product attributes or components. Analysis of such features for the identification and prediction of the sentiment of the document is called as feature based sentiment analysis.

IV. ISSUES IN SENTIMENT ANALYSIS

Sentiment analysis being a popular research topic in the machine learning domain, has some of the issues while analyzing the text. As stated by Bing Liu, in his book "Sentiment Analysis and Opinion Mining" [6], an opinion may be positive for certain group of people, but the same opinion may be considered as negative by other set of people. Here are some of the known issues of sentiment analysis,

1. Interrogative sentences may be analyzed incorrectly because the interrogative sentence may include some of the keywords which will result in either positive or negative opinion, even if the sentence is objective.
2. Sarcastic sentences are the most challenging part to be handled in case of sentiment analysis. Most of the sarcastic sentences will be having more positive polarity, but even though they look positive, the meaning which they convey will be in a negative manner. Being a human it is easy to catch the sarcastic nature of the sentence by understanding the intonation used. But the system will fail sometimes while analyzing the sentiment of the statement.
3. There are some sentences without any sentimental words in it, but will convey either positive or negative sentiment. For example, "This is a seven star hotel", in this statement we have no sentiment words, but the phrase 'seven star' conveys that it is a very good hotel.
4. Use of substitute word for a usually practiced and normal word can give inappropriate result while analyzing the text. For example, the word "good" is usually spelled as 'gud' in most of the chat conversations.
5. Perception, is a concept where it deals with how a person perceives the sentence. As stated earlier a sentence can be positive for one and the same can be negative for the other. For example, "Dollar price is increasing with respect to Indian rupee".
6. Some of the sentences or in particular reviews can be a spam while analyzing and can give an opposite result.

V. APPLICATIONS OF SENTIMENT ANALYSIS

The sentiment analysis can be applied in many different domains. The major use of sentiment analysis is to analyze the texts which are available on the social networking sites, where people share their views on particular product or a topic. Companies use the sentiment analysis tools in order to analyze the market, by analyzing the customer feedbacks. Now-a-days even politicians are appointing data analysts to analyze the political party related texts available on the news reporting websites and also in social networking sites. Some of the major applications of the sentiment analysis are as stated below,

1. In Review-Related Websites.
2. The sentiment analysis can be used as a Sub-Component Technology.
3. It can be used in Business and Government Intelligence.
4. Sentiment analysis can be used across Different Domains.

VI. SYSTEM FRAMEWORK

We have proposed a model (Figure 4) which takes the text as the input. The text can be any review or any text which can be used for the sentimental analysis. There are two major layers in the proposed model, data processing layer and sentiment analysis layer. The data processing layer deals with data collection, data pre-processing and data mining, while the sentiment analysis layer is meant for the actual analysis of the text and to present the result. More details will be introduced in the following sections.

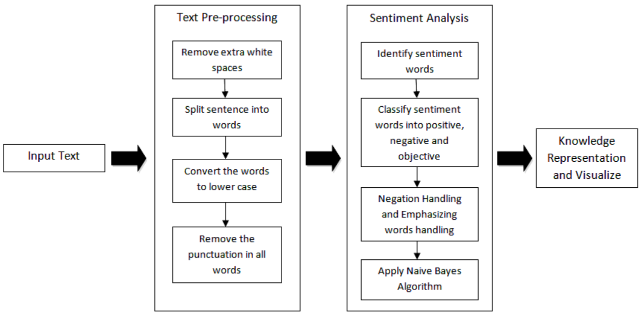

Input Text
Text Pre-processing
Remove extra white spaces
Split sentence into words
Convert the words to lower case
Remove the punctuation in all words
Sentiment Analysis
Identify sentiment words
Classify sentiment words into positive, negative and objective
Negation Handling and Emphasizing words handling
Apply Naive Bayes Algorithm
Knowledge Representation and Visualize

Figure 4: Proposed System Architecture

A. *Input and Text pre-processing*

The input to the system can be any desired text to be analyzed. The text can be either a sentence or paragraphs (reviews or any other text). The text is first loaded into the system and the text is pre-processed. The pre-processing phase includes the removal of the while spaces in the text entered, splitting the sentence into words, later these words are converted to a common case i.e. lower case in this context. In order to analyze or process the words, there should be no punctuations, so the punctuations are removed from the words.

B. *Text processing and sentiment analysis*

Before entering into the text processing phase, we should have the set of positive and negative words in-order to match the words in the sentence with the words available in the data dictionary. We have used the information provided by SentiWordNet (SWN) [15], in our research work to get the actual polarity (positive, negative and objective) of the words available in English language. The first phase in the text processing and sentiment analysis phase is the sentiment word identification. Wherein which the polarity of the words are added up, which results in three possible polarity scores positive, negative and objective polarity scores. This gives a way for classification of the words into positive, negative or objective.

C. *Naive Bayes Classifier*

Naive Bayes classifier is a simple yet powerful probabilistic model based on Bayes rule and independence assumption [2]. This model simplifies conditional independence assumption. That is given a class (positive or negative) the words are conditionally independent of each other [14]. In our case, the maximum likelihood probability of a word belonging to a particular class is given by the expression:

$$P(y_i|c) = \frac{\text{Count of } y_i \text{ in text of class } c}{\text{Total no of words in text of class } c} \tag{1}$$

Where y_i are the individual words in the text. But, according to Bayes rule the probability of a particular document belonging to a class c_i is given by,

$$P(c_i|d) = \frac{P(d|c_i) * P(c_i)}{P(d)} \quad (2)$$

If we use the simplifying conditional independence assumption, that given a class (positive or negative), the words are conditionally independent of each other. Due to this simplifying assumption the model is termed as "Naive".

$$P(c_i|d) = \frac{(\prod P(y_i|c_j)) * P(c_j)}{P(d)} \quad (3)$$

This classifier outputs the class with the maximum posterior probability.

D. *Negation Handling*

Negation handling is the most important phase in the sentiment analysis [18]. It can also be stated as a major problem during the process of sentiment analysis [14]. The usage of the negation words like not or the words ending with n't (Example: Can't, won't and so on) will change the positive sentiment into a negative sentiment and vice versa. For example, "This phone is not good.", this statement will be considered as a positive sentiment if the negation words are not handled. But if the negation words are handled, the proper sentiment will be fetched from the system. Here is the pseudo code of the negation handling phase in our system.

```
PSEUDO CODE:
negated_negative = False
negated_positive = False
for each word in text
    if word before positive_word is "not" or "n't word"
        negated_positive = True
    if word before negative_word is "not" or "n't  word"
        negated_negative = True
```

This process of negation handling resulted in a significant improvement in the accuracy of sentiment analysis. But later we discovered that only negation handling was not sufficient enough to get proper and more accurate results, so we decided to include another mechanism i.e. Emphasizing words handling.

E. Emphasizing Words Handling

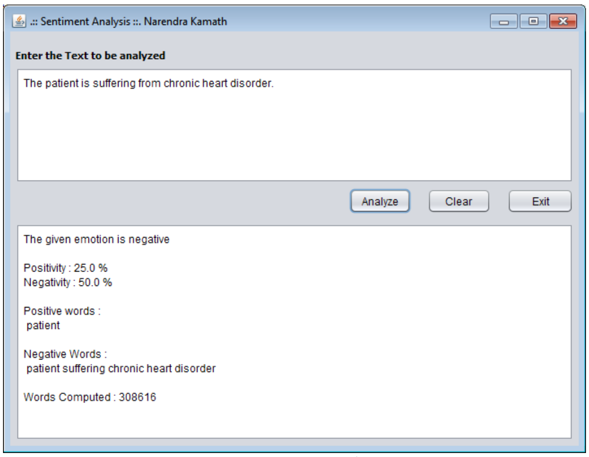
As of now, we have discussed about increasing the accuracy of the sentiment analysis system by using negation words handling. But there are some words which strengthen the sentiment value either positive or negative polarity value. The use of words like very, high, increases and so on will emphasize the words which are used before or after these words. For example, "Unemployment increases the crime rate in the city". After we handled these emphasizing words, there was even more accurate results.

F. Feature Selection

Feature selection is the process of eliminating the redundant set of features, while selecting only those features from the dataset, which are most useful or which are most relevant. The use of bigrams and trigrams, will result in a problem resulting in the increment in the features [14]. Most of these features are redundant and noisy in nature. To reduce the number of available features and to eliminate the noise, we can follow a basic filtering method [4].

VII. RESULTS

Experiments are conducted with various categories of text. The algorithm and the application was developed in java, since it is a widely used and a platform independent language. The test of the system was done from the basic level of sentences till the complex reviews found on some review sites. The system was tested with some complex sentences, which included sarcasm statements, medical statements and the sentences which included both positive as well as negative sentiments but resulting in a single sentiment. The results are found accurate as of now.



Enter the Text to be analyzed

The patient is suffering from chronic heart disorder.

The given emotion is negative

Positivity : 25.0 %
Negativity : 50.0 %

Positive words :
patient

Negative Words :
patient suffering chronic heart disorder

Words Computed : 308616

Figure 5: Experimented result of Entered Dataset1

The entered text is "The patient is suffering from chronic heart disorder". This sentence is analyzed accurately as a negative sentiment, the positivity and the negativity in the sentence is calculated and displayed, with the positive and negative words present in the sentence. Here is the computed result with the visualization of the same.

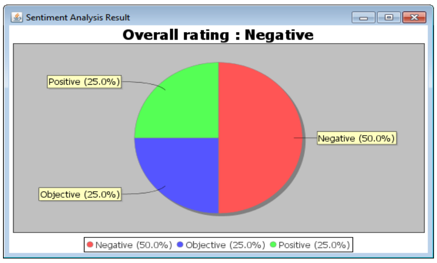
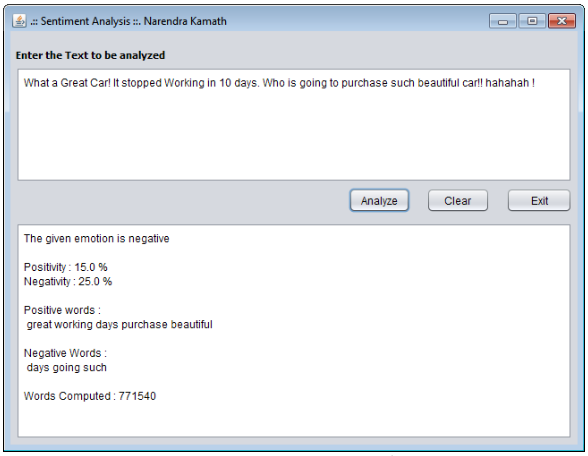


Figure 6: Visualized result of the Dataset1

By consideration of the sarcastic emotion and the emphasizing words, we have the experimental results of two texts.



Enter the Text to be analyzed

What a Great Car! It stopped Working in 10 days. Who is going to purchase such beautiful car!! hahahah !

The given emotion is negative

Positivity : 15.0 %
Negativity : 25.0 %

Positive words :
great working days purchase beautiful

Negative Words :
days going such

Words Computed : 771540

Figure 7: Experimented result of Dataset2

The above experimented result of a sarcastic text was taken from a post which was posted on one of the social networking sites. Here we can see that even though the positive words are more, the result obtained is accurate to be a negative sentiment, since it is a sarcastic sentence. Here is the visualization of the same result.

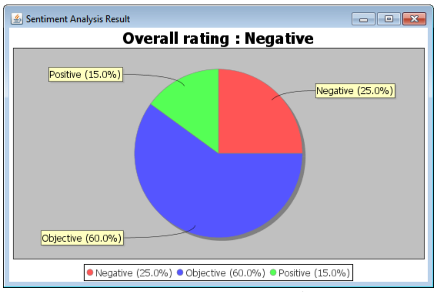
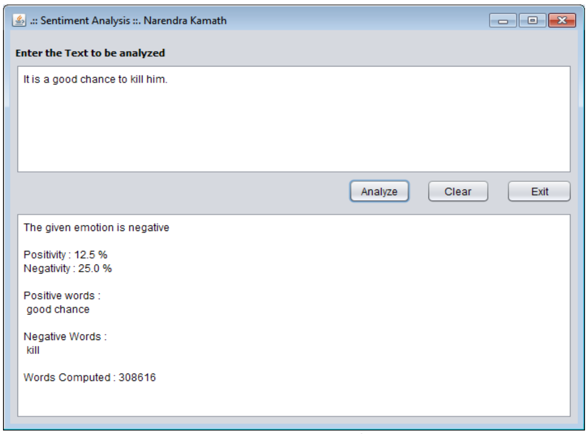


Figure 8: Visualized result of Dataset2

Some sentences will be having many positive words, but those positive words may refer to the negative word which is used in later part of the sentence. By emphasizing the negative words by using the positive words, will increase the degree of negativity in the sentence. So, we experimented with such kind of sentence. The exprimented sentence was, "It is a good chance to kill him". This sentence has the positive words 'good' and 'chance' but the phrase 'good chance' refers to the negative word 'kill' thus the result expected was negative sentiment. The system gave the same and accurate result as negative.



Enter the Text to be analyzed

It is a good chance to kill him.

The given emotion is negative

Positivity : 12.5 %
Negativity : 25.0 %

Positive words :
good chance

Negative Words :
kill

Words Computed : 308616

Figure 9: Experimented result of entered Dataset3

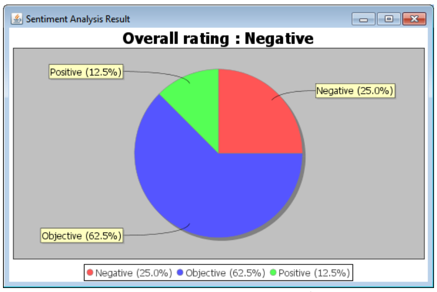


Figure 10: Visualized result of Dataset3

VIII. CONCLUSION

Sentiment analysis is a very important aspect in data analytics and the advanced text processing. Each word in a given text is related to the adjacent or any other related words. Merely by looking at the occurrences of the positive or negative words in a text, we cannot conclude with a sentiment which it results in. As of now we have seen that by implementing Naive Bayes classifier, Negation Handling, Emphasizing words Handling and by feature selection, we have got highly significant improvement in the sentiment classification.

IX. REFERENCES

[1] Liu , Lei Zhang "A SURVEY OF OPINION MINING AND SENTIMENT ANALYSIS ".

[2] Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24 2014.

[3] Bo Pang and Lillian Lee Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[4] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, Clustering Product Features for Opinion Mining, WSDM'11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493-1/11/02...$10.00

[5] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proc. of the ACL, pages 271–278. ACL, 2004.

[6] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[7] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

[8] Abhishek Tiwari, Kshitij Pathak, Upasana tiwari, Rupam das, Opinion Polarity Detection in Blog Comments from Blog Rss Feed by Modified TF-IDF Algorithm, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 1, Jan-Feb 2012, pp. 412-416

[9] Hemalatha, Dr. G. P Saradhi Varma, Dr. A.Govardhan, "Sentiment Analysis Tool using Machine Learning Algorithms", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 2, Issue 2, March – April 2013 ISSN 2278-6856

[10] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar,"OPINION MINING AND ANALYSIS: A SURVEY", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013.

[11] "Sentiment Classification using Machine Learning Techniques", Shivakumar Vaithyanathan IBM Almaden Research Center.

[12] "Opinion Mining and Sentiment Analysis", Bo Pang, Computer Science Department, Cornell University, Ithaca, NY 14853, USA.

[13] Rohini K. Srihari,"OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction", Department of Computer Science & Engineering University of New York at Buffalo.

[14] "Fast and accurate sentiment classification using an enhanced Naive Bayes model.",Vivek Narayanan, Department of Electronics Engineering, Indian Institute of Technology (BHU), Varanasi, India.

[15] http://sentiwordnet.isti.cnr.it/

[16] "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10) (May 2010).

[17] "A Novel Approach for Sentiment Analysis and Opinion Mining", Dr. Ritu Sindhu, Ravendra Ratan Singh Jandail, Rakesh Ranjan Kumar, International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 4, April 2014).

[18] Das, Sanjiv, and Mike Chen. "Yahoo! for Amazon: Sentiment parsing from small talk on the web." EFA 2001 Barcelona Meetings. 2001.