ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 5, May 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Machine Learning Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets

G. Vaitheeswaran[1]
Research Scholar, Department of Computer Science
St. Joseph's College (Autonomous),
Tiruchirappalli – India

L. Arockiam[2]
Associate Professor, Department of Computer Science
St. Joseph's College (Autonomous),
Tiruchirappalli – India

Abstract: ***One of the most promising application is analysis of sentiments in social networks. Sentiment analysis deals with determining the sentiment orientation—positive, negative, or neutral—of a text. In recent times, it attracts abundant interest on both academia and industry due to its valuable applications. In this paper it is examined, how classifiers work while doing opinion mining over Twitter data. Reducing the data size using the feature selection method produces better accuracy and increase the computational space. The feature selection method plays a vital role in increasing the accuracy of sentiment analysis. The selected features for the research work are unigrams, negation words, emoticons, stemming and retweet count. The retweet count plays a major role in sharing others' opinion. The ranking method is used to select the top most and relevant features. The best-ranking method for the text mining is Zipfs' law and is used to rank the selected features. The proposed Senti_Classi approach is experimented with Naïve Bayes, Support Vector Machines and Maximum Entropy. The 10 cross-fold validation method is used for training and testing the classifiers. This paper presents the best machine learning approach to sentiment analysis on tweets.***

Keywords: sentiment analysis; machine learning; Zipfs' law; support vector machine; naïve Bayes; maximum entropy.

I. INTRODUCTION

Sentiment analysis (SA) is the field of study that analyses peoples' opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes [1]. It is also called as opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they all are under the umbrella of sentiment analysis or opinion mining. The sentiment analysis can be performed using the three main approaches. They are as follows:

- *Lexicon Based Approach:* It is also called as dictionary based approach and relies on a lexicon or dictionary of words with pre-calculated polarity. The lexicon can be generated using manual dictionary or from some other available corpus.
- *Machine Learning Based Approach:* It requires a corpus containing wide number of tagged examples. The machine learning tasks involve with statistical learning and probabilistic methods.
- *Hybrid Approach:* The combination of lexicon and machine learning method are known as hybrid approach.

Sentiment analysis task can be transformed into classification task, so machine learning techniques can be used for sentiment analysis. In this article, existing research study of sentiment analysis includes supervised learning classifiers such as Naïve Bayes, Support Vector Machines, and Maximum Entropy are analyzed.

The framework of the current article is as follows: In the section 'background study', a brief review on types and processing of machine learning have been discussed. The existing works on sentiment analysis using machine learning

approaches have been provided in the 'related works' section. In the 'objective' section, the motivation and objective of the proposed work are explained in detail. The proposed approach has been explained in the section 'Senti_Classi Approach'. The obtained results are analyzed and discussed under the section 'results and discussions'. Following the previous section, the overall summary of this article is concluded in the end section of this article.

II. BACKGROUND STUDY

In 1959, Arthur Samuel, defined machine learning as a, "Field of study that gives computers the ability to learn without being explicitly programmed". It is closely related to computational statistics. It is a method of data analysis that programs, analytical model building. In simple term, the machine learning is like how the human brain learns from life experiences and from open guidelines.

According to Tom M. Mitchell, the machine learning can be defined as, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [2].

Here are a few widely exposed examples of machine learning application of social media:

- Knowing what customers are saying about the product, sports, politics, disaster, etc. on Twitter.
- Online recommendation offers like those from Amazon and Netflix.
- Fraud detection.

A. Types of Machine Learning

The machine learning tasks are typically classified into four broad categories as follows:

- Supervised learning: It is the machine learning task of inferring a function from labeled data.
- Unsupervised learning: It is the machine learning task of inferring a function to describe hidden structures from unlabeled data.
- Semi-supervised learning: It is a class of supervised learning tasks and techniques that also make use of unlabeled data for training. Typically it is a combination of supervised and unsupervised learning method (a small amount of labeled data with a large amount of unlabelled data).
- Reinforcement learning: It is an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take action in an environment so as to maximize some notion of cumulative reward.

B. Processing of Machine Learning for Sentiment Classification

Figure 1 represents the process of the machine learning tasks for sentiment classification. The process of machine learning for sentiment classification can be broken into the following steps:

1. Data pre-processing.
2. Feature selection.
3. Learning an Algorithm.
4. Model Evaluation.

1. Data pre-processing

The most common phrase "garbage in, garbage out" is predominantly relevant to any kind of data mining. The data pre-processing is to remove the noisy and redundant data from the raw data sources. The raw data from excel, access, text files, social media data, etc., (gathering past data) forms the foundation for the future learning. Any data of good quality lead the analytical process to a good result. To get the quality data, pre-processing is the important step in the data analytics for fixing issues such as missing data and treatment of outliers. The product of data pre-processing is the final training set.

2. Feature Selection

Features are text attributes that are valuable for capturing the patterns hidden in the text. The most popular features used in machine learning classification are the frequency of n-grams extracted during the preprocessing step. In the frequency-based representation the number of occurrences of a particular n-gram is used. In case if the text length varies greatly, it might be important to use Term Frequency (TF) and Inverse Document Frequency (IDF) measures. However, in short messages like tweets, words are unlikely to repeat within one occurrence. Apart from the n-grams, additional features can be created to improve the overall quality of text classification. The most common features that are used for this purpose include [3]:

- Number of words with positive/negative sentiment
- Number of negations
- Number of exclamation marks and emoticons
- Number of different parts-of-speech in a text (for example, number of nouns, adjectives, verbs, etc.,)

Since the main features of text classifier are N-grams. The dimensionality of the feature space grows proportionally to the size of the dataset. The growth of the feature space makes the most cases computationally infeasible to calculate all the features of a sample. Many features are redundant or irrelevant and do not significantly improve the results. Feature selection is the process of identifying a subset of features that have the highest predictive power. This step is crucial for the classification process, since elimination of irrelevant and redundant features reduces the size of feature space. And this process also helps to increase the speed of the algorithm to improve the quality of classification. The most commonly used feature selection methods are listed below:

- Chi-square test [4]
- Mutual Information [5]
- Information Gain [3]
- Term frequency – Inverse document frequency (tf-idf) [6]
- Zipfs' law [7]

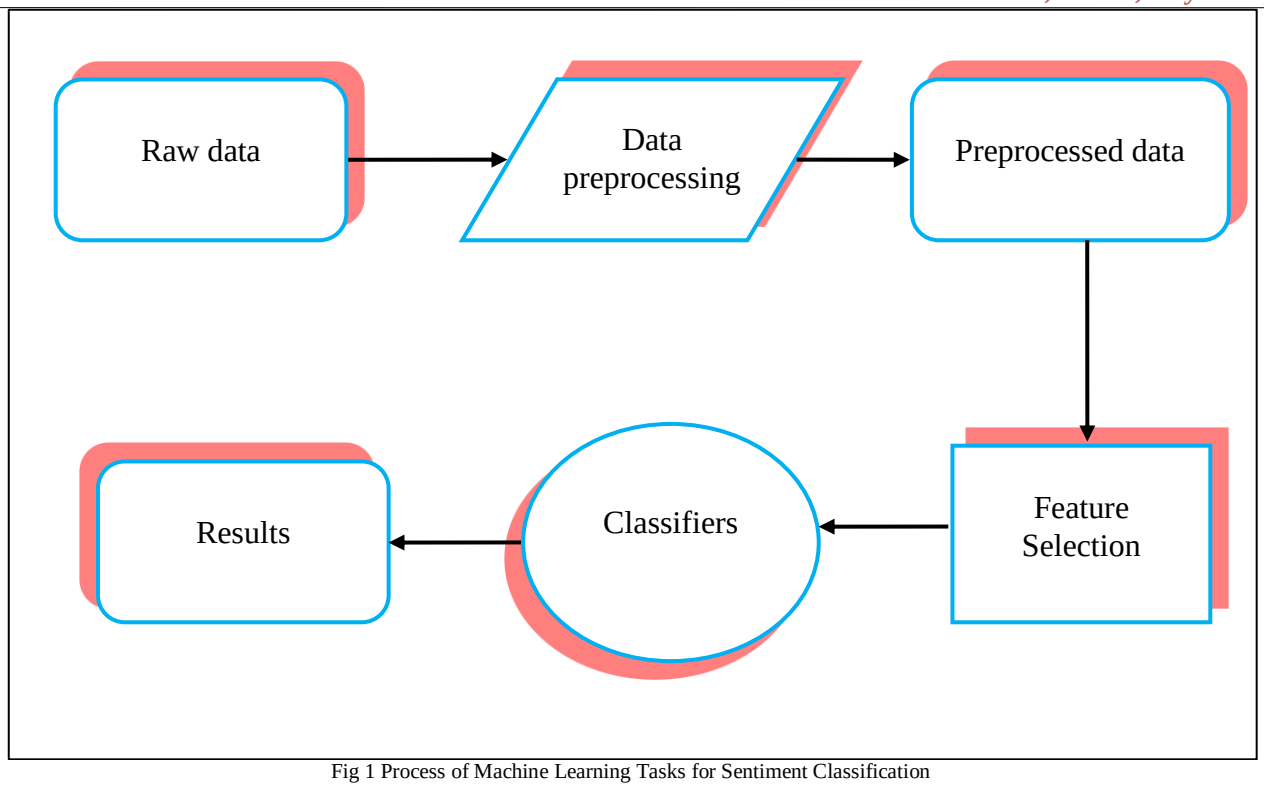

Fig 1 Process of Machine Learning Tasks for Sentiment Classification

3. Learning an Algorithm

This step involves choosing the appropriate algorithm and representation of data in the form of a model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites). The first part (training data) is used for developing the model. The second part (test data), is used as a reference.

4. Model Evaluation

To test the accuracy, the second part of the data, i.e., hold out or test-data is used. This step determines the precision in the choice of algorithm based on the outcome. A better test to check the accuracy of the model is to see its performance on data which was not at all used during the model build.

III. Related Work

Reshma et al. [8] had proposed the automatic document classification for Tamil documents. The most popular supervised learning algorithms such as Maximum Entropy (ME), Conditional Random Fields (CRF) and Support Vector Machines (SVM) were used to achieve sensible accuracy in both the sentence and document level classification. The Tamil newspaper Dinakaran dataset obtained from the EMILLE corpus had been utilized to experiment the ability of Machine Learning algorithms in Tamil domain classification. The N-gram features had been used to learning the model. The work had been experimented with enormous amount of dataset; the problem of space complexity had not raised with ME and CRF algorithms. But SVM had shown the degraded performance towards space complexity problem while doing classification at sentence level.

Yogesh et al. [7], had focused on classifying tweets based on the sentiments expressed them into three categories namely positive, negative and neutral. The author investigated the space of Twitter Sentiment analysis by using two-step classifier and negation detection. In addition, the author claimed that an efficient sentiment analyzer is considered to be a must in the era of big data, where a great number of electronic communications are a major bottleneck. Major difficulties had been found in handling the tweets due to their limited size, and the cryptic style of writing that makes them difficult to understand for better analysis. Two classifiers were designed based on Naive-Bayes and Maximum Entropy classifiers, and their accuracies were compared on different feature sets. The most popular frequency based feature selection method Zipfs' law was used to select the most relevant features.

Olga Kolchyna et al. [3] stated that, how machine learning approach tackles the problem as a text classification task by employing two supervised classifiers – Naïve Bayes and Support Vector Machines. The most relevant features were selected using "Information Gain" evaluation algorithm and a "Ranker" search method from WEKA package. The features considered for the work were as follows:

- N-grams
- Elongated words number
- Emoticons
- Negation
- POS
- Punctuation marks
- Negative tokens number
- Positive tokens number

The accuracy was compared with the Sem-Eval 2014 competition and shown better results.

IV. OBJECTIVE

From the above related works, most of the work had been carried out by using the supervised learning algorithms such as Naïve Bayes, Support Vector Machines and Maximum Entropy to find the accuracy of tweets. Most of the existing work had been carried out using tf-idf method for feature selection. For the unigram and bigram model, Zipfs' law based feature selection produced better results.

This motivated to build a new model based on selecting different features using the Zipfs' law to enhance the accuracy. The objectives of the proposed approach are:

- To enhance the accuracy using the best-supervised learning model for the collected datasets.
  - To find the topmost relevant features to reduce the space complexity.
  - To find the best supervised learning model for the collected datasets.
  - To find the precision, recall, f-score and accuracy for positive, negative and neutral tweets.

V. PROPOSED APPROACH

Senti_Classi Approach

The concept of the proposed Senti_Classi approach is to evaluate the sentiment knowledge on tweets using machine based approach. Figure 2 exhibits the overall procedure for the proposed work. The labeled tweets obtained from the Senti_Lexi approach [8] are used to evaluate the proposed approach.

Approach: Senti_Classi.
Find the best learning model for the collected labeled tweets and to enhance the accuracy.
Input: Labeled tweets

- Find the best learning model.
- Enhance the accuracy of sentiment classificationa

Method:

1. Preprocessed datasets
2. Apply Zipfs' law to select the relevant features
3. Cross validation method (10 fold method)
4. Learning the model
5. Evaluate the model

Fig 2 Senti_Classi Approach

Step 1 of Senti_Classi appraoch is to load the preprocessed data into the classifiers model. In step 2, the frequency based feature selection method is used to select the most relevant features based on the ranking method. For this method, the classifier model used the Zipfs' law. Instead of using the training and testing dataset separately, the cross validation method is used. The 10 crossfolds are set to learn the model (step 3). Step 4 applies the classifier model (SVM, NB and ME) to find the best models for the collected datasets. Step 5 evaluates the best model based on the accuracy.

Feature Selection

Feature selection is the process of selecting a subset of relevant features which have the highest predictive power. This step is vital for the classification process to eliminate irrelevant and redundant features. Due to such elimination, the size of feature space is greatly reduced and the computational speed increased and in turn improve the quality of classification. New features can be created which are expected to improve the overall quality of text classification.

The possible number of n-grams (where, n>4) exponentially increases the size of n-gram and increase the training set size. This exponential growth makes the classifier model infeasible to calculate all the features of a sample in a limited amount of time. Many features are redundant or irrelevant and do not meaningfully impact the classification of results. To avoid these constraints, the unigrams and bigrams models are used to train the model.

This section describes the features that are generated and used to train the classifier. The selected features of this research work are as follows:

- ***Unigrams***: The presence of positive, negative and neutral words in the tweets.
- ***Bigrams***: The negation words are tokenized along with the opinionated word and it is considered as the features.
- ***Emoticons***: Presence of positive, negative and neutral emoticons in the tweets.
- ***Stemming***: The elongated words are stemmed and tokenized. The stemmed words are considered as the features in order to increase the accuracy.
- ***Retweet count***: In the existing work, the retweet count was eliminated and not consider as the feature. The retweet count is considered as another feature in the proposed Senti_Classi approach. Since the positive, negative and neutral tweets posted by people are shared by other users, to show their state of mind towards the target. This shows the target contains strong polarity. To reduce the features space, the unigram and bigram features are extracted along with the presence of retweet count. This produces more accuracy.

Frequency based feature selection

Frequency-based feature selection is the process of selecting the terms that are most frequently occurs in the class. Frequency can be either defined as a document frequency (the number of documents in the class 'c' that contain the term 't') or

as a collection frequency (the number of tokens of 't' that occur in documents in 'c'). When many thousands of features are selected, the frequency-based feature selection often does well. Thus, if somewhat suboptimal accuracy is acceptable, then frequency-based feature selection can be a good alternative to more complex methods.

Zipfs' law

Zipf's law establishes the relationship between the frequency of any word in the text and its rank. It is one of the most well-known and widely used laws in linguistic informatics. Conventionally, Zipf's law has been employed to reduce the index size and improve the processing speed of data retrieval system and text stratification.

Zipfs' law frequency based ranking

Zipf's law is a law about the frequency distribution of words in a language (or in a collection that is large enough so that it is representative of the language). Let r be the rank of a word, Prob(r) be the probability of a word at rank r. The concept of the law is to find the rank of the word based on the frequency. The equation 1 represents the definition of the Zipfs' law.

By definition

$$Prob(r) = \frac{Freq(r)}{N} \qquad \text{.....Eq (1)}$$

where

Freq(r) = the number of times the word at rank r appears in the collection.

N = total number of words in the collection (not number of unique words).

Selected top most features

The table 1 displays the top 10 features produced by the frequency based ranking in the preprocessed tweets. The most occurred word is "Love" and the ":)" is the emoticon expressed by the most number of people.

Table 1 Top Features based on Frequency Ranking

S. No	Features
1.	:)
2.	Love
3.	Happy
4.	Amazing
5.	:(
6.	Crash
7.	like
8.	Not_worth
9.	adorable
10.	Won't

Classifier Models

The cross validation method is used for training and testing the model. The value for cross validation is set to 10. The classifier models used for the proposed approach are as follows:

- Naïve Bayes,
- Support Vector Machines and
- Maximum Entropy.

Naïve Bayes (NB)

A Naïve Bayes classifier applies Bayes' theorem in an attempt to suggest possible classes for any given text. To do this, it needs a number of previously classified documents of the same type. The theorem is as follows:

Bayes theorem provides a way of calculating posterior probability P(A|B) from P(A), P(B) and P(B|A). Look at the equation 2 below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{..... Eq (2)}$$

Above,

- P(A|B) is the posterior probability of class (A, target) given predictor (B, attributes).
- P(A) is the prior probability of class.
- P(B|A) is the likelihood which is the probability of predictor given class.
- P(B) is the prior probability of predictor.

To find the probability of a tweet for positive, negative and neutral classes, the equations 3, 4 and 5 states the classification of positive, negative and neutral tweets respectively.

$$P(positive|tweet) = \frac{P(tweet|positive)P(positive)}{P(tweet)} \quad \text{..... Eq (3)}$$

$$P(negative|tweet) = \frac{P(tweet|negative)P(negative)}{P(tweet)} \quad \text{..... Eq (4)}$$

$$P(neutral|tweet) = \frac{P(tweet|neutral)P(neutral)}{P(tweet)} \quad \text{..... Eq (5)}$$

Support Vector Machines (SVM)

The support vector machine is a supervised learning algorithm that works as follows:

- It uses a nonlinear mapping to transform the original training data into a higher dimension.
- Within this new dimension, it searches for the linear optimal separating hyperlane.
- With an appropriate nonlinear mapping to high dimension, data from two classes can be separated by hyperlane.
- The SVM finds this hyperlane by using the support vectors ("essential training tuples") and margins (defined by the support vectors).
- The presence and absence of features are represented using the binary value 1 and 0 respectively.

The SVM is a binary classifier, i.e., the class labels can only take two values: +1 or -1. The fig 3 represents the visualization of two-class SVM.

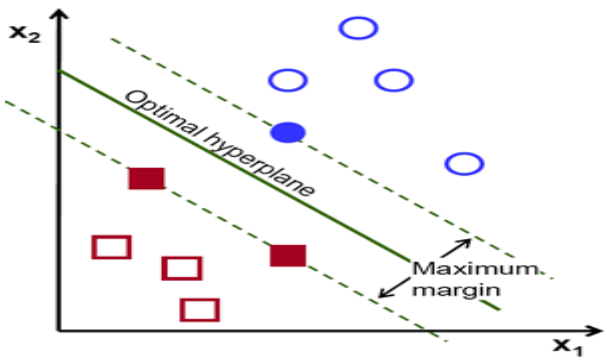

Fig 3 The 2-D SVM Based Model on Linearly Separable Data

In this case, the labeled data sets contain three classes, namely positive, negative and neutral, and these come under the 'three-class' problem. To build binary classifiers for multi-class SVM, the class labels are distinguished as follows:

(i) between one of the labels and the rest (*one-versus-all*) or

(ii) between every pair of classes (*one-versus-one*).

The one-versus-one based classifier is used to learn the model. To evaluate the model, the binary Sequential Minimal Optimization (SMO) is applied by using the following one-versus-one classifier:

- positive, neutral
- positive, negative
- neutral, negative

The kernel function used for the evaluation is, Polynomial kernel of degree d. The equation 6 states the function of polynomial kernel.

$$k(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d \quad \text{…….. Eq (6)}$$

Maximum Entropy (ME)

Maximum Entropy (ME) classifier is an another supervised learning technique, and that has proven effective in a number of text mining applications. It is a probabilistic classifier which belongs to the class of exponential models. It is also known as a conditional exponential classifier which converts labeled feature sets to vectors using encoding. In some cases, the ME classifiers outperform the Naïve Bayes at standard text classification. The equation 7 represents the exponential form of ME [Pan, 08].

$$P_{ME}(c|d) = \frac{1}{Z(d)} * exp\ (\sum_I \lambda_{i,c} F_{i,c}(d,c)) \quad \text{…….. Eq (7)}$$

Where, Z (d) is a normalization function. $F_{i,c}$ is a feature/class function for feature f_i and class c, as in Eq. 8.

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d)>0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases}$$

…..... Eq (8)

VI. Results and Discussions

The table 2 shows the comparative results obtained from the Support Vector Machines (SVM), Naïve Bayes (NB) and Maximum Entropy (ME) models. These results are analyzed and interpreted from the following figure 4.

Table 2 Comparative Table of SVM, NB and ME

Measures	SVM (%)	NB (%)	ME (%)
Precision	84.70	83.70	81.10
Recall	84.30	82.30	80.30
F-Score	83.90	82.99	80.70
Accuracy	85.71	84.30	82.14

The figure 4 represents the value comparison of supervised machine learning techniques such as Support Vector Machines, Naïve Bayes and Maximum Entropy with each other. These techniques are evaluated using the common measures such as precision, recall, F-Score and accuracy. The X- axis in the chart diagram represents the classifier algorithms. The Y-axis represents the percentage of the common measures. The cost of time taken to learn the model Support Vector Machines is

greater than the Naïve Bayes and Maximum Entropy. But the Support Vector Machines yields higher accuracy when compared with Naïve Bayes and Maximum Entropy.

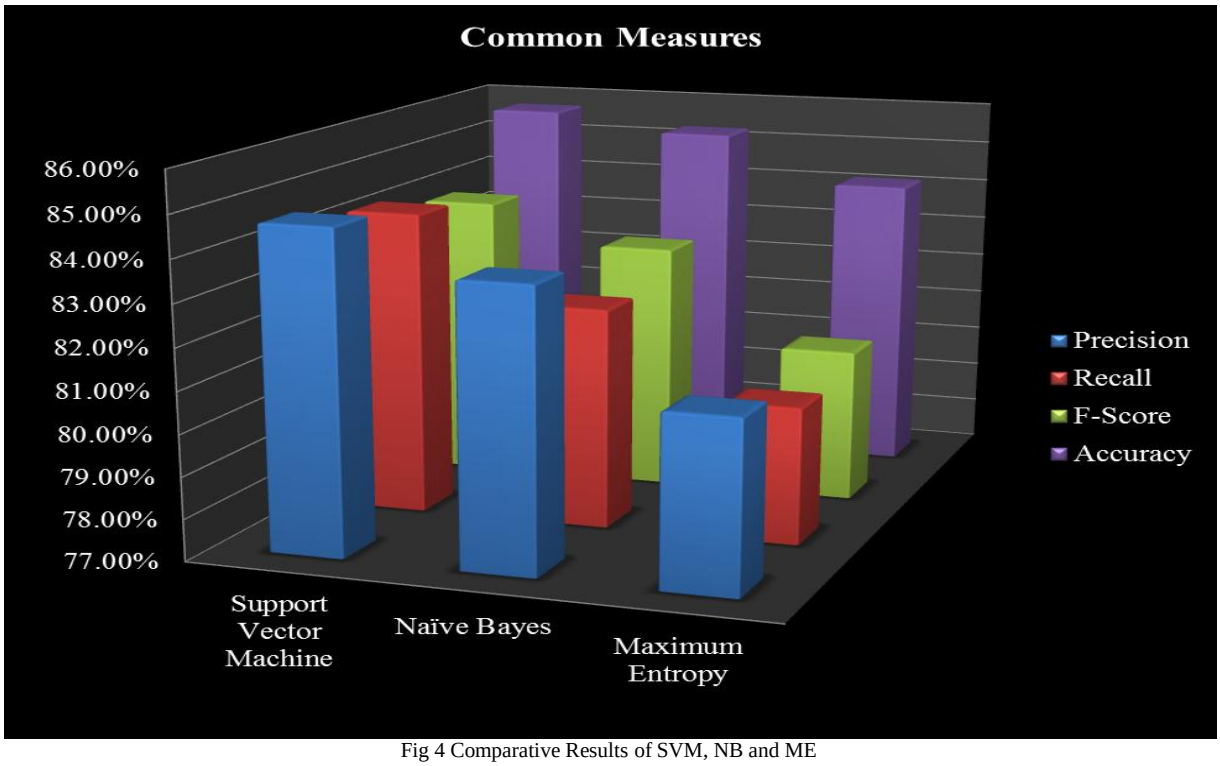

Fig 4 Comparative Results of SVM, NB and ME

VII. CONCLUSION

This article clearly shows the machine learning based approach for sentiment analysis. An attempt made to explain the process of machine learning tasks in the background study. The supervised machine learning techniques such as Naïve Bayes, Maximum Entropy and Support Vector Machines have been discussed. The labeled datasets obtained from the Senti_Lexi approach are filtered by adding the retweet count as another feature to the Senti_Classi approach. This feature increases the computational space and reduces the probability sparseness. Features are selected on the basis of feature based ranking method, by using the familiar Zipfs' law.

The proposed approach is based on the unigrams and bigrams model which produced better accuracy than the existing work. The overall performance of the three classifier models produces accuracy more than percentage of 80. The SVM based classifier model produced better accuracy than the maximum entropy and Naïve Bayes. Many of the existing works on sentiment analysis are based on bag-of-words and unigrams, which do not capture the context dependent word and is essential for sentiment analysis. Concentrating on this context dependent word produces more accuracy.

References

1. Bing Liu. "Sentiment Analysis And Opinion Mining", Morgan and Claypool publishers, 2012.
2. Tom M. Mitchell, "Machine Learning", published by McGraw-Hill Science/Engineering/Math, ISBN: 0070428077, 1997, pp. 2.
3. Olga Kolchyna, Tharsis T. P. Souza, Philip C. Treleaven and Tomaso Aste1, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination", arXiv: 1507.00955v3 [cs.CL], 2015.
4. L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories, 2011.
5. Walaa Medhat, Ahmed Hassan and Hoda Korashy, "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, Vol 5, 2014, pp. 1093-1113.
6. Vo Ngoc Phu and Phan Thi Tuoi, "Sentiment Classification using Enhanced Contextual Valence Shifters", International Conference on Asian Language Processing, IEEE, pp. 224 - 229.
7. Yogesh Garg and Niladri Chatterjee, "Sentiment Analysis of Twitter Feeds", Third International Conference, BDA 2014, New Delhi, India, Spirnger Proceedings, 2014, pp 33-52.

8. Reshma U., Barathi Ganesh H. B., Anand Kumar M. and Soman K. P, "Supervised Methods for Domain Classification of Tamil Documents", ARPN Journal of Engineering and Applied Sciences, Vol 10, No. 8, 2015, pp. 3702-7.
9. G. Vaitheeswaran and L. Arockiam, "A Novel Lexicon Based Approach to Enhance the Accuracy of Sentiment Analysis on Big Data", International Journal of Emerging Research in Management and Technology (IJERMT), Volume 5, Issue 2, January 2016.

