Sentimental Analysis of Political Tweets In Twitter Using Machine Learning Algorithm

K. Ragaventhari[1], Mrs. E. Siva Shankari[2]
[1] Dept of computer science and engineering
[2]Asst. Professor, Dept of computer science and engineering
[1, 2] Government college of Engineering, Tirunelveli.

Abstract- *Sentiment analysis is an area of research that analyzes opinions, sentiments, evaluations, attitudes, and emotions from a written text. The feelings and emotions of people have a key effect in our day by day process. Especially social media is used to analyze sentiment of peoples, product, etc,. Now a days social media play a vital role in this matter and it has received much attention. Public and personal opinion a few big variety of subjects are expressed and spread continually via numerous social media. Especially Twitter is one of the social media that is gaining popularity. It offers effective way to analyze people's perspective towards political parties and events. Twitter sentiment analysis is an application of sentiment analysis of data in Twitter (tweets), to extract sentiments conveyed by the user. The tweet format is very small which generates a whole new dimension of problem like use of slang, abbreviations etc. In this work, the general user tweets are reviewed from the election point of view and the tweets are studied from the user view of political events. The proposed system develops a classification model to identify the political orientation of twitter users based on the tweet content. Naïve Bayes classifier is used in this process of analysis. On evaluation, Naïve Bayes outperforms SVM by 5%.*

Keywords- Social media, Twitter data, Machine learning, Naïve Bayes, SVM.

I. INTRODUCTION

Now a days, social media plays a vital in politics and it is used to predict an election result. Especially the younger generations are keen interest to known about the politics because of social media. Social media is used to analyze a peoples opinion because peoples are keen interest to post their opinion on social media about products, trending topics, political news, politicians, movies etc… This plays a large unstructured information for data mining. With over 100 million register users, the most popular social media platforms such as Facebook, Youtube, WeChat, Instagram, QZone, Weibo, Twitter, Tumbir, Telegram, Baidu Tieba, LinkedIn, LINE, Snapchat, Pinterest, Viber, etc.

Twitter is a website for micro blogging and social networking, on which users post and communicate messages known as tweets. Tweets were originally limited to 140 characters but on November 7, 2017, this limit was expanded to 280 for all languages except Chinese, Japanese and Korean. Registered users can post, like, and retweet for tweets, but unregistered users can only read them. Users can access Twiter through interface, Short Message Service (SMS) or Mobile Apps. Twitter is a platform that is widely used by individuals to express their opinions and show feelings on various occasions. Analysis of sentiment is the automated process that analyzes text data and classifies views as negative, optimistic, or neutral. [1] It is a process of deriving sentiment of a particular statement or sentence and technique of classification that draws opinion from the tweets and formulates a sentiment. Sentiment are subjective to the topic of interest. We are required to formulate that what kind of features will decide for the sentiment it embodies. [1] The approaches to sensitivity analysis can be broadly categorized into two classes. They are lexicon based and machine learning based. Lexicon based approach is unsupervised as it proposes to evaluate opinions, Whereas machine learning approach involves use of feature extraction and training the model using feature set and some dataset.

Machine learning algorithm are very helpful to classify and predict whether a particular document have positive and negative sentiment. Machine learning is categorized into two types known as supervised and unsupervised machine learning algorithm. Supervised learning algorithm used as labeled dataset where each document of training set is labeled with appropriate sentiment, whereas, unsupervised learning include unlabelled dataset where text is not labeled with appropriate sentiments.

This paper primarily focus on supervised machine learning techniques on a unlabelled dataset. Evaluation of the sensitivity is usually carried out on three levels, namely the level of sentence, document level and dimension. First level, aims at classifying the entire document or topic as positive or negative. Second level, consider the polarity of a document's individual sentence while third level classification determines

the aspects of a corpus, and then the polarity is determined for each document with respect to the aspects obtained for market exploration. Twitter sentiment analysis using advanced text mining technique to analyze the sentiment of the text in the form of positive, negative and neutral.

Machine learning approach is rather simpler than knowledgebase approach. Actually, there are different machine learning techniques that are used to classify data i.e., they are naïve Bayes classifier, support vector machine, decision tree, random forest, neural networks etc.

II. RELATED WORK

Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani [1] said what is sentiment analysis and twitter sentiment analysis, how to analyze a sentiment on twitter, which techniques are used to classify the sentiments, how to use the python language for sentiment analysis.

Kaviya Supala, Narasinga Rao [2] did sentiment analysis on twitter dataset by downloading the tweets on developing a Twitter API. Classification is done using the Naïve Bayes algorithm and its performance is increased by pre-processing the tweets. The final results shows the classification of text in their required classes.

Huma Parveen and Prof. Shikha Pandey [3] did sentiment analysis on movie dataset by downloading the tweets on developing a Twitter API. They have used the Hadoop framework for processing the dataset. Classification is done using the Naïve Bayes algorithm and its performance is increased by pre-processing the tweets. The final results shows the classification of text in their required classes

Neethu M S and Rajasree R [4] performs analysis on tweets based on some specific domain using different machine learning techniques.They tried to focus on problems that are faced during the identification of emotional keywords from multiple keywords and difficulty in handling misspellings and slang words. So a feature vector is created whose accuracy is tested using naïve bayes, SVM, maximum entropy and ensemble classifiers.

Bac Le and Huy Nguyen [5] built a model to analyze the sentiment on Twitter using machine learning techniques by applying effective feature set and enhances the accuracy i.e., bigram, unigram and object-oriented features. The classification of tweets is done using two algorithms i.e., Naïve Bayes classifier and Support vector machines(SVM) whose accuracies are tested by calculating precision, recall and fscore and also shows same accuracy.

III. METHODOLOGY

Existing Paper: The existing system analyzes User Tweets using Hashtags and Keywords. The system collects tweets using HashTags which are nothing but the popular personalities/parties. The existing system collects tweets and perform volume analysis using map reduce. However trend analysis involves collecting tweets of popular or trending party/candidate and performing analysis to bifurcate the positive and negative tweets for the party/candidate. This helps the party/candidate to act accordingly to improve their reputation and help user to actually make a clear opinion about any party/candidate. This is conducted in 3 phases. Initially connect with tweeter and download the tweets. Then load these tweets on HDFS for further analysis and finally perform actual analysis or volume analysis or Trend Analysis or Sentiment analysis to gain information.

Proposed System: The proposed system analyzes sentiment with machine learning techniques on twitter data about political tweets. In order to perform sentiment analysis, collect data from the desired source (Twitter). After the collection of dataset, it is divided into training and testing sets. The training set is the main aspect upon which the result depends. This data undergoes various steps of pre-processing which makes it more machine sensible than its previous form. The data pre-processing is a very important step as it decides the efficiency of the other steps. It involves syntactical correction of the tweets as desired. The steps involved should aim for creating the info more computer readable so as to scale back ambiguity in feature extraction. After the data pre-processing, some features are extracted with the help of bag-of-words. A feature is a piece of information that can be used as a characteristic which can assist in solving the problem (like prediction). The quality and quantity of features is very important as they are important for results generated by Term Frequency Inverse Document Frequency (TF-IDF). After feature extraction, the sentiments are classified with the help of machine learning techniques such Naïve Bayes and SVM. In this work, Python SciKit Learn library is used as the tool for implementing the proposed work as it increases the efficiency of machine learning supervised. This entire workflow is depicted in Figure 3.1

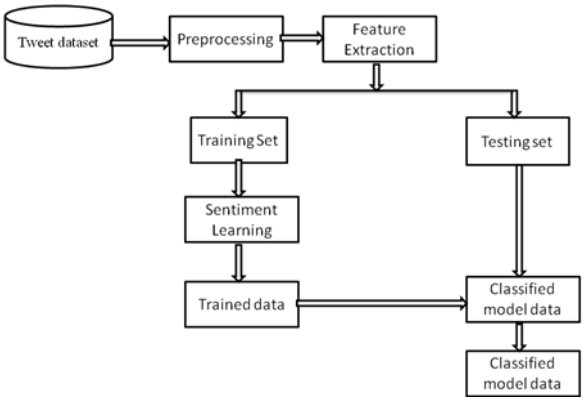

Fig: 3.1 System Architecture

IV. MODULE DESCRIPTION

Implementation of twitter sentiment analysis for political events using four modules.

4.1 Modules

- Dataset Collection
- Data Pre-processing
- Feature Extraction
- Sentiment Classification

4.1.1 Dataset Collection

In this proposed work twitter dataset is required. The dataset was gathered from kaggle site. The downloaded dataset is a csv format. It contains many fields but the auspol.csv is used because the data are labeled as a positive and negative. It also contains positive data range and negative data range and also mention which label is best fit for the tweet. The dataset contain 45269 tweets extracted using he twitter API keywords. It can be used for detect sentiment on political tweets.

4.1.2 Data preprocessing

After the collection of twitter dataset about political events, that data was splitted into training and testing set. The training dataset is going to preprocessing step. The preprocessing of the data is a very important step as it decides the efficiency of the other steps. It involves syntactical correction of the tweets as desired because the real world data is often incomplete, inconsistent, and lacking in certain behavior or trends, and is likely to contain many errors. The following steps involved should aim for making the data more machine readable in order to reduce ambiguity in feature extraction.

Step 1: Removal of re-tweets: In the tweet text, we can usually see that every sentence contains a reference that is called retweet. Because it is repeated through a lot of tweets and it doesn't give us any information about sentiment that's why we can remove them.

Step 2: Converting upper case to lower case: In order to bring all tweets to a consistent form. By performing this, we will assure that further transformations and classification tasks won't suffer from non-consistency or case sensitive issues in our data.

Step 3: Stop words removal: Stop words are function words that are high frequently present across all tweets. There is no need for analyzing them because they do not provide useful information. We can obtain a list of these words from NLTK stop words funtions (for example and, or, still etc.).

Step 4: Twitter feature removal: User names and URLs are not important from the perspective of future processing, hence their presence is futile. All usernames and URLs are converted to generic tags or removed .

Step 5: Special character and digits removal: The special character will generate tokens with a high frequency that will cloud our analysis and digits do not contain any sentiment, so it is important to remove them from tweet text.

Step 6: Tokenization: It is the process of converting text into tokens before transforming it into vectors. It is also easier to filter out unnecessary tokens. For example, a document into paragraph or sentence into words. In this case we are tokenizing the reviews into words.

Step 7: Stemming: Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming algorithm attempt to automatically remove suffixes in order to find the "root word" or stem of a given word.

Step 8: Lemmatization: it's the method of grouping together the various inflected sorts of a word in order that they are often analysed as one item. It is similar to stemming but it brings context to the words.

4.1.3 Feature Extraction

After the data preprocessing, we need to extract the feature from preprocessed data. The feature extraction is used here to identify key features in the data for coding by learning from the coding of the original data set to derive new ones. A technique for tongue processing that extracts the

words(features) utilized in a sentence, document, website, etc,. and classify them by frequency use.

It can better and improve the accuracy of learning algorithm and shorten the time. It also used to irrelevant or redundant features. In this process, we need to use TF-IDF technique.

TF-IDF(Term Frequency-Inverse document frequency) uses all the tokens in the dataset as vocabulary, Frequency of occurrence of a token from vocabulary in each document consists of the term frequency and number of documents during which token occurs determines the Inverse document frequency. If a token occurs frequently in a document that token will have high TF but if that token occurs to frequently in majority of documents then it reduces the IDF.

Term Frequency (TF): The number of times a word appears in a sentence divded by the total number of words in the sentence. Every sentence has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Inverse Data Frequency (IDF): The log of the number of sentence divided by the number of sentence that contain the word *w*. Inverse data frequency determines the weight of rare words across all sentence in the corpus.

$$idf(w) = \log\left(\frac{N}{df_t}\right)$$

4.1.4 Sentiment Analysis

After the feature extraction, we need to classify the data into positive, negative and neutral labels. In this work, Naïve Bayes classification algorithm is used.

4.1.4.1 Naïve Bayes Algorithm

Naïve Bayes classification technique based on Bayes theorem, which is the foundation of deductive reasoning, which focuses on determining the probability of an event occurring based on prior knowledge of conditions that might be related to the event. The Naïve Bayes Classifier brings the power of this theorem to Machine Learning building a very simple yet powerful classifier.

$$P(H|E) = (P(E|H) * P(H)) / P(E)$$

Where,

P(H|E) is the probability of hypothesis H given the event E, a posterior probability.
P(E|H) is the probability of event E given that the hypothesis H is true.
P(H) is the probability of hypothesis H being true, or prior probability of H.
P(E) is the probability of the event occurring.

4.1.4.2. Support Vector Machine

A Support Vector Machine (SVM) is a type of supervised machine learning classification algorithm. SVM differs from the opposite classification algorithms within the way that it chooses the choice boundary that maximizes the space from the closest data points of all the classes.

An SVM doesn't merely find a choice boundary, it finds the foremost optimal decision boundary. In this proposed system was implemented by python. In python, we need use Scikit Learn library for implementing the SVM. To divided the data into training and testing data, Scikit-Learn contains the svm library, which contains built-in classes for different SVM algorithm. Since we are going to perform a classification task, use the support vector classifier class, which is written as a svc in the Scikit-Learn's svm library. In SVC class, the fit method is called to train the algorithm on the training data, which is passed as a parameter to the fit method.

After the training data, to make predictions, the predict method of the svc class is used. In the case of a simple SVM we simply set this parameter as "linear" since simple SVMs can only classify linearly seperable data.

$$g(X)= w^{T}\emptyset(X)+ b$$

X is feature vector, 'w' is weights of vector and 'b' is bias vector is the non-linear mapping from input space to high dimensional feature space.

V. RESULT AND DISCUSSION

5.1 Data Prerprocessing

After the collection of dataset (Table 5.1) read that csv file into python platform The data preprocessing is to removal of digits and twitter features from tweets. After that the data is to be tokenized, that token words are splited each word in document, then the stop words are removed and convert uppercase to lowercase from token words. Then the stemming is processed using porter stemmer algorithm. After stemming, lemmatization was applied on stemmed words. Table 5.2 show the preprocessed tweets. The data pre-

processing was split the dataset into training set and testing set using Naïve Bayes algorithm

Table 5.1 tweets before preprocessing

s.no	Tweets
1	Id like to see Andy Burnham as the leader of the Labour party.
2	ITV News poll: Labour could be largest party as they close gap on Tories
3	Every more all these cannot be wrong don't be blinded by the fake party's vote
4	is not perfect, by any means, but they are the party with the responsibility to face any problem Britain faces
5	Mirror: "The Labour Party would gain the most if people could vote online" < That translates as "Labour voters are lazier than others"
6	You don't even need to be 'political' to understand that the labour party aren't going to improve our conditions realistically

Table 5.2 tweets after Preprocessing

s.no	Preprocessed Tweets
1	Id like see Andy Burnham leader Labour party.
2	ITV News poll: Labour could large party close gap Tories
3	Every wrong don't be blinded by the fake party's vote
4	Perfect means, party responsible face problem Britain faces
5	Mirror Labour Party gain most people vote online translate Labour voter lazy
6	You dont even need political understand labour party go improve our condition realist

5.2 Feature Extraction

The feature extraction is used to extract feature from the dataset, which is implemented by TF-IDF technique, TF is used to identify the number of times a word appear in a sentence divide by the total number of words in a sentence. Then IDF is used to determines the weight of the all words in sentence.

5.3 Sentiment Analysis

In this proposed work, Naïve Bayes and SVM algorithm is used to analyze a sentiment for political events. We have to classify the label such as positive, negative and neutral using Naïve Bayes classifier and SVM. This proposed system was implemented by python In Naïve Bayes, It also divide the data into training and testing sets using train_test_split from sklearn.model_selection library in python. To train the Naïve Bayes on the training data. Scikit-Learn contains the naïve bayes library, which contains built-in classes for different Naïve Bayes algorithms. Since to perform a classification task, we will use the support vector classifier class, which is written as MultinominalNB in the Scikit-Learn's naïve bayes library. In SVM, we have divided the data into training and testing sets. To train our SVM on the training data. Scikit-Learn contains the svm library, which contains built-in classes for various SVM algorithms. Since to perform a classification task, we will use the support vector classifier class, which is written as SVC in the Scikit-Learn's svm library.

Table 5.3 Labelled Tweets

s.no	Tweets	Labels
1	Id like see Andy Burnham leader Labour party.	positive
2	ITV News poll: Labour could large party close gap Tories	negative
3	Every wrong don't be blinded by the fake party's vote	negative
4	Perfect means, party responsible face problem Britain faces	neutral
5	Mirror Labour Party gain most people vote online translate Labour voter lazy	positive
6	You dont even need political understand labour party go improve our condition realist	positive

5.4 Performance Analysis

Accuracy - Accuracy is that the most intuitive performance measure and it's simply a ratio of correctly predicted observation to the entire observations. The accuracy is a great measure but only when we have symmetric datasets where values of false positive and false negatives are almost same.

Therefore, we have to look at other parameters to evaluate the performance of our model.

$$Accuracy = (TP+TN)/(TP+FP+FN+TN)$$

Precision - Precision is that the ratio of correctly predicted positive observations to the entire predicted positive observations. High precision are related to the low false positive rate.

$$Precision = TP/(TP+FP)$$

Recall (Sensitivity) - Recall is that the ratio of correctly predicted positive observations to the all observations in actual Recall = TP/(TP+FN)

F1 score - F1 Score is that the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives under consideration. Intuitively it's not as easy to know as accuracy, but F1 is typically more useful than accuracy, especially if you've got an uneven class distribution. Accuracy works best if false positives and false negatives have similar values or range. If the value of false positives and false negatives are very different, it's better to seem at both Precision and Recall.

$$F1\ Score = 2*(Recall * Precision) / (Recall + Precision)$$

Table 5.4 Table for twitter sentiment analysis

S,no		SVM	Naïve Bayes
1.	Accuracy	0.754755	0.800000
2.	Precision	0.75415	0.800000
3	Recall	0.752794	0.857143
4	F1Score	0.753251	0.7532251

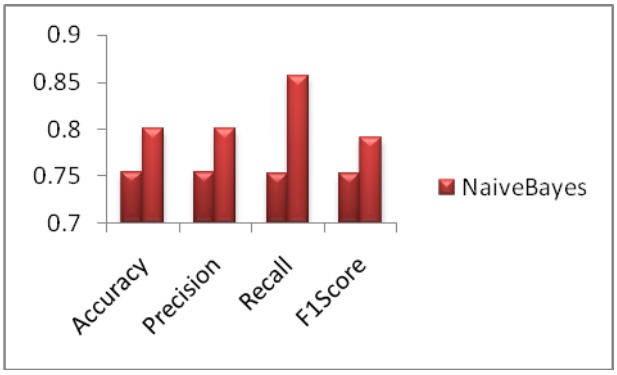

Fig 5.2 Graph for evaluation of twitter sentiment analysis

Fig 5.2 show the evaluation of a performance measure such as Accuracy, Precision, Recall and F1score for our proposed system.

VI. CONCLUSION AND FUTURE WORK

6.1 Conclusion

Sentiment Analysis is the process of analyzing the general opinion of the audience. It may be a reaction to a piece of news, movie or any tweet about some matter under discussion. Generally, such reactions are taken from social media and clubbed into a file to be analyzed through NLP. It is widely used for getting insights from social media comments, survey responses, and product reviews, and making data-driven decisions. Twitter Sentiment Analysis is a text mining technique for analyzing the underlying sentiment of a text message.

This work has been carried out to enhance the accuracy of prediction by Navie Bayes classifier and Support Vector Machine classifier. Sentiment analysis, a branch of digital analytics aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of the document. This work is investigated whether sentiment analysis of public mood derived from large scale Twitter feeds can be used to identify how many supporter and hater for political events. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy. It contains a lots of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. In SVM, the accuracy of the prediction is 75% . In Navie Bayes, the accuracy of the prediction is 80 %. In comparison, Navie Bayes produce a high level of accuracy.

6.2 Future Enhancement

Twitter Sentiment Analysis is used to analyze a opinion for audience in any topic. It is expressed as positive, negative (or) neutral. No algorithm can give a 100% accuracy (or) prediction on sentiment analysis. The future work of this research can employ other Word Vectorization techniques such as Count Vectorizer and Word2Vec and also other classification algorithm like Neural Networks.

REFERENCES

[1] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi 2017, ' Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python', International

Journal of Computer Applications, vol. 165, no. 9, pp. 0975 – 8887.

[2] Kavya Suppala, Narasinga Rao 2019, 'Sentiment Analysis Using Naïve Bayes Classifier', International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 8, pp. 2278-3075.

[3] Huma Parveen, Prof. Shikha Pandey 2016, 'Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm', International Conference on Applied and Theoretical Computing and Communication Technology, pp. 416-419.

[4] M. S. Neethu and R. Rajasree 2013, 'Sentiment analysis in twitter using machine learning techniques,' Fourth International Conference on Computing, Communications and Networking Technologies, pp. 1-5.

[5] Le B., Nguyen H. (2015), 'Twitter Sentiment Analysis Using Machine Learning Techniques'. In: Le Thi H., Nguyen N., Do T. (eds) Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, vol 358. Springer, Cham

[6] Varsha Sahayak, Vijaya Shete and Apashabi Pathan 2015, 'Sentiment Analysis on Twitter Data', pp. 2349-2163.

[7] Peiman Barnaghi, John G. Breslin and Parsa Ghaffari 2016, 'Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment', 2016 IEEE Second International Conference on Big Data Computing Service and Applications.

[8] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau 2011, 'Sentiment Analysis of Twitter Data' Proceedings of the Workshop on Language in Social Media (LSM 2011).

[9] Aliza Sarlan, Chayanit Nadam and Shuib Basri 2014, 'Twitter Sentiment Analysis', International Conference on Information Technology and Multimedia (ICIMU), pp. 18 – 20.

[10] Gayatri P.Wani and Nilesh V.Alone 2015, 'Analysis of Indian Election Using Twitter' International Journal of Computer Application, vol. 22, pp. 0975-8887.

[11] Gayatri Wani , Nilesh Alone, "A Survey on Impact of Social Media on Election System".

[12] Aparna.U.R and Shaiju Paul 2016, 'Feature Selection and Extraction in Data Mining', Online International Conference on Green Engineering and Technologies.

[13] M. Pechenizkiy, S. Puuronen, A. Tsymbal, 'Feature Extraction For Classification In The Data Mining Progress', International Journal "Information Theories & Application".

[14] Amardeep Kaur and Jagroop Kaur 2018, 'Sentimental Analysis and Method', International Journal Of Computer Science and Engineering, vol. 6, no. 7.

[15] Jo Williams, Susan J. Chinn and James Suleiman 2014, 'The Value Of Twitter For Sports Fans', Journal of Direct, Data and Digital Marketing Practice 16.