ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse103912020.pdf
https://doi.org/10.30534/ijatcse/2020/103912020

Twitter Sentiment Analysis: A Political View

Joylin Priya Pinto[1], Vijaya Murari T.[2], Soumya Kelur[3]

[1]Department of Computer Science and Engineering, NMAMIT, Nitte, Karnataka, India, priyahirgan@gmail.com
[2]Department of Computer Science and Engineering, NMAMIT, Nitte, Karnataka, India, vijayamurari.t@nitte.edu.in
[3]Department of Computer Science and Engineering, YIT, Moodabidri, Karnataka, India, pkelursoumya16@gmail.com

ABSTRACT

Twitter is a famous micro-blogging service for tracking public mood with respect to an object or entity. It involves an application of sentiment analysis on naturally written language to extract sentiments conveyed by the users. Twitter sentiment analysis is demanding. Because of the raw nature of text, limited size, use of slang words, abbreviations, tweet processing task became quite complicated. But, one cannot ignore the twitter data, as it hold sentiments towards variety of topics. This research aims to analyse political orientation of twitter users as positive or negative. Different machine learning algorithms are adopted to identify the user point of view on Ayodhya issue. Then the efficiencies of these built models are contrasted with one another to discover the best machine learning algorithm on text categorisation. The result is analysed based on the measurement metrics namely accuracy, precision, recall, F1-score measures for each sentiment class.

Key words: Sentiment Analysis, Ayodhya, Feature Extraction, Random Forest, SVM, KNN, NB, Logistic Regression

1. INTRODUCTION

Twitter has risen as a major social-networking site. With such enormous gathering of people, it is ceaselessly pulling in clients to pass their conclusions and point of view about any issue. Because of this reason, Twitter is utilized as a data source by various domains. On Twitter, clients are permitted to express their perspectives as tweets, which is constrained to just 140 characters. In past few years, lot of research is carried on Twitter data in order to analyse people sentiments. The major aspect of performing opinion mining on tweets is fundamentally to distinguish the tweets into various target classes precisely.

Twitter data analysis is challenging, as stated earlier. Justification can be given as followed:

- Tweet size limitation: with the restriction of 140 characters, limited clarifications will be conveyed, which does not highlight important tweet features.
- Slang words usage: Slang words are not exactly the English words which makes the strategy obsolete in view of the transformative utilization of slangs.
- Twitter features: Twitter permit the usage of hashtags, client references and URL links. These need extra tweet handling than typical English words.
- User variety: People convey their point of views in many ways, some use variety of languages in the middle, while others may utilize rehashed words or symbols to pass on a feeling.

Every one of these issues are need to be handled in the pre-processing stage.

1.1. Twitter Sentiment Analysis Using Python

Python is one of the most powerful programming language. It is popular for its code intelligibility and conservative line of codes. It focuses on indentation and uses white space to delimit the blocks. Python gives an enormous standard libraries for various applications such as natural language handling, artificial intelligence, and data analytics and so on.

1.2 Natural Language Tool Kit (NLTK)

NLTK is considered as a standard python library, which gives the platform for content handling and categorization. Natural Language Processing (NLP) is an important method to examine, comprehend, and get significance from human understandable language in a brilliant and valuable manner. It is the best technique to perform Sentiment Analysis. Human language is plainly spoken. Understanding plain language is not only to comprehend the spoken words, but also to analyse the thoughts and how they are connected together to produce proper meaning. Despite language being one of the clearest thing for human to learn, the ambiguity of language is the most difficult issue for machines to master.

NLP methods are generally relied on machine learning techniques. Rather than hand-coding huge set of rules, NLP can depend on machine learning to automatically learn these rules by analysing a set of examples such as a large corpus, gathering of sentences and making a statistical inference. As and when the more data is analysed, the model will become more accurate. Operations like tokenizing,

labelling, text manipulation can be carried out by the utilization of NLTK package.

2. LITERATURE REVIEW

Research is going on tremendously in the area of sentiment analysis. Some of the important reviewed papers are discussed below:

Gupta et al. [1] presented a novel method which deals with location-wise tweets and meanwhile compared the working efficiency based on accuracy. Author implemented Naïve Bayes and Support Vector Machine classifiers in order to group the twitter data into positive and negative tweets. At that point, the locations are put into categories and sentiment mapping is done which helped in analysing the opinions of individual Indian states separately. Important features were considered for classification such as latitude, longitude, kilometres and number of tweets to distinguish the opinions based on region. Importance of dataset pre-processing is also demonstrated in the paper. However the author recognized state-wise people reactions to demonetization, because of population polarization, the general feeling of the citizens couldn't be caught.

Gautama et al. [2] focused on three supervised learning algorithms namely SVM, Naïve Bayes and maximum entropy for the twitter data classification. Tweets are classified based on the expressed sentiments. Obtained results are contrasted with respect to performance measures such as: accuracy, precision and recall. A semantic analysis is driven from WordNet sentiment dictionary subsequent to training and categorisation. Consequently, a relative measurement is done on the classification through supervised machine learning techniques and semantic analysis.

Tsapatsoulis et al. [3] tried to label tweets manually for the training dataset preparation. Then these manually annotated tokens are tested to check whether this dataset can form an index of terms that can be utilized for creating viable tweet classification models. Then the manually labelled index of terms are compared with various automatically extracted feature sets in the categorisation of tweets using machine learning framework. Author used three distinct algorithms for the justification purpose. Three approaches were identified namely lexicon approach, machine learning strategy and the social strategy. From the observation, it is found that the explicitly identified tokens gave better performance accuracy than the other feature extraction methods. However, the feature set combination is not considered in the result observed by the author.

Taiwo Kolajo et. al. [4] proposed an advanced method for pre-processing the social media tweets. He used Nigerian tweets as source text. With the help of Twitter API, tweets subjected to politics are collected. More concentration was given on local knowledge such as slang words are identified during pre-processing stage. Ambiguous words are resolved with the help of Lesk algorithm on context basis. Models are built using SVM, Multi-layer Perception (MLP) and Convolutional Neural Network(CNN). Obtained result proved that CNN performs well than SVM and MLP. Built model got an accuracy of 99.17% while classifying Naija tweets.

Parveen et al. [5] demonstrated on HDFS and MapReduce architecture. While implementing, initially the dataset is pre-processed. After that, a supervised machine learning technique named Naïve Bayes is enforced on the dataset for the classification. A trained SentiWordNet dictionary is used while training the classifier. In this review, two methodologies have been used to implement Naïve Bayes technique. First method is based on map stage where SentiWordNet dictionary content is read from a file and transformed to Hashmap table for key-value based polarity extraction of words which makes the process faster. Second methodology consists of reduce stage, where the overall sentiment polarity of individual tweet is gathered. Then the identified polarity is grouped into five classes as strongly positive, positive, strongly negative, negative and neutral. However, the sentiment classification is divided among 5 different classes, the implementation part is carried out assistance of just a single classifier.

Abdelwahab et al. [6] studied the impact of training dataset size variation on SVM and Naive Bayes classifiers. Author also analysed the impact of shifting the training dataset size on the learning curves to absorb information of both SVM and Naïve Bayes classifiers when utilized in twitter opinion examination. Along with this, the effectiveness of the training dataset on various ensemble types is tested. When two ensemble methods are compared with each other, it is found that AND combined result of Support Vector Machine and Naïve Bayes classifiers is good enough than the ensemble 2 method where SVM and Naïve Bayes algorithms are OR combined. However, consolidating the aftereffects of classifiers brought about equivocal outcomes for the correlation between Naive Bayes and SVM.

Neethu et al. [7] considered two unique methods for extricating the sentiments from dataset; symbolic approach and the machine learning strategy. Symbolic approach utilises the lexical resources which are available in unsupervised opinion categorisation. In machine learning strategy, various techniques such as Random Forest, Maximum Entropy, Support Vector Machine are implemented to categorise user opinions based on extracted features. The author identified that Machine Learning strategy is more efficient and simpler than Symbolic approach.

Sahni et al. [8] utilized the subjectivity of sentences to categorise dataset. Normally an absolutely objective sentence does not pass on any assumption. Therefore, just subjective sentences are used here in the training set. In the implementation stage, different natural language processing libraries are adapted to identify the subjective sentences. Thereafter, the various classifiers are applied on the pre-processed dataset. Distinctive feature removal methods namely n-grams, POS tagging etc are utilised to extract significant features.

Shohreh Haddadan et al. [9] presented an approach to mine sentiments from political debates. He considered US presidential campaign debates for sentiment classification. Proposed method is carried out using two steps i.e. sentences that express argument are detected in the first step. In the second step, classification was done at the sentence level. Author used SVM and Neural Network for classification. Fallacy detection was an important factor which was introduced in the proposed work.

Anjali Bhavan et al. [10] focused on the UK parliament debates classification. He used publicly available dataset along with manually annotated sentiment labels. Graph based features are used for analysis. Sentiment classification models are generated in two parts. One is based on manually annotated labels and the other one is based on government dataset. After obtaining classification models, it is observed that the graph based features performed well in contrast with non-graph based features.

3. METHODOLOGY

Figure 1 shows the Architectural diagram of the implemented sentiment analysis framework.

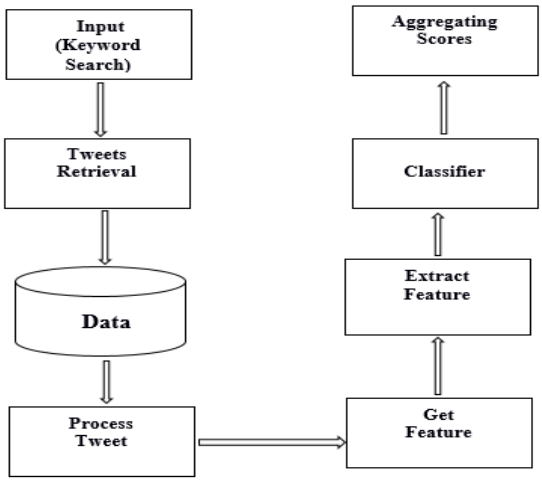

Figure 1: Proposed System Architecture

3.1 Data Collection

Tweet gathering is the process of tweet collection related to specific hashtags. Tweepy - a Twitter Streaming API is used in the proposed work for the collection of twitter data. In order to scrape tweets from Twitter, one need to enrol for the Twitter Application Development account. Once the user is registered with the developer account, keys and access tokens will be provided, which are mandatory to access twitter tweets. Consumer Key and Consumer Secret keys, similarly Access token and Access Token Secret keys are need to be copied into the code, using which tweets can be collected dynamically, each time when the code is executed.

The tweets are gathered here, in view of Ayodhya related hashtags for an ideal timeframe of analysis from December 2018 to April 2019. In supervised machine learning approach, efficiency of the model depends on the training dataset. The ratio of dataset into training and testing set plays vital role in obtaining more efficient classification model. The training set is the primary viewpoint whereupon the outcomes depend.

In the proposed work, while scraping data from Twitter API, bag of words method is used in the Vader sentiment analyser to score the tweets and based on the sentiment scores tweets are labelled to create training dataset.

3.2 Tweet Pre-processing

Tweet cleaning phase is a significant stage upon which the efficiency of the processes depend, which are need to be carried out in later stages. It includes grammatical correction of the tweets as needed. The processes included should make the information more machine readable so that the uncertainty in feature extraction could be reduced. Pre-processing of tweets involve the below mentioned steps:

- Case conversion: In case of case sensitive analysis, there is a possibility of considering two same words as different because of its sentence case. There is a necessity to perform case conversion from upper case to lower in order to do effective sentiment analysis. Without which, the model may give wrong prediction.
- Stop words extraction: Stop words does not pass on any sentiments. Therefore, they can be ignored in the pre-processing phase.
- Twitter feature removal: During feature handling, presence of User names and URLs can be ignored as they does not convey any feeling.
- Stemming: Replacing words with their root words, decrease various kinds of words with comparative implications. This aides in dimensionality reduction of the feature set.
- Special character and numbers evacuation: Numbers and special symbols does hold opinions. Most of the times, symbols are mixed with words, subsequently evacuation of these help in effective tweet analysis.
- Emoticon Dictionary: Using a lexicon to process emoji's simplifies the task of sentiment analysis.
- Removing hashtags: With the usage of regular expressions, hashtags can be expelled.
- Removing rehashed tweets.

3.3 Feature Removal

Feature removal/extraction is the first step while training a classifier with machine learning approach. It is a way to convert each text into number format to represent in the vector form. Without which the classifier cannot perform text classification. To get better classification model, it is necessary to have good number of qualitative and quantitate features. For removing features, various procedures are accessible in the present day. Most commonly used feature extraction technique is 'Bag Of Words' which works on the basis of occurrence of words in a text document or sentence i.e. frequency of words is considered as feature vector. Initially, only the opinion bearing words are extracted from

the sentence which always exhibits subjective opinions, then the machine learning classifier is fed with training data that consists of extracted feature set pairs and the classification model is produced. When the new model is trained enough with data samples, it can perform more accurate prediction. The same feature extractor is utilized in the transformation of unseen text data in to feature set pairs which can be fed into the classification model to make predictions.

While performing classification on huge amount of data, machine learning strategy is normally more accurate. The classifiers like Random Forest cannot use "Bag-of-Words" feature extraction method directly. At this point, Term Frequency-Inverse Document Frequency is more efficient. With TF-IDF, words are given weight-TF-IDF measures relevance, not frequency. That is, word counts are replaced with TF-IDF scores across the whole dataset. First, TF-IDF measures the number of times that words appear in a given document i.e. "term frequency". But because words such as "and" or "the" appear frequently in all documents, those must be systematically discounted. That's the inverse-document frequency part. The more frequently a word appears in documents, the less valuable that word is as a signal to differentiate any given document. That's intended to leave only the frequent AND distinctive words as markers.

Scikit-learn provides Vectorizer to translate input documents into feature vectors. Library function TfidfVectorizer(), is used to provide parameters for the kind of features required, by indicating the number of acceptable features.

3.4 Sentiment classifiers

In this step, learning model is imported, based on which target prediction should be done. Scikit-learn is a python based machine learning library that facilitates machine learning techniques which are best suited to perform sentiment analysis. To use machine learning classifiers, initially NumPy should be installed, which is the basic package for scientific computations in Python. Then the desired model is imported from scikit-learn. Once the model is trained, the same can be for testing later.

From scikit-learn tool, following Machine learning classifiers are imported and text classification has been done on those of built models.

- Support Vector Machine

SVM is considered as a best machine learning technique, which works well on both linear and non-linear classification problems. While performing classification, it searches for an isolating hyperplane. Whenever there is a maximum possible distance between two target classes, then that is treated as best hyperplane. Support vector machine works well on text categorisation problems. Sometimes, it is very difficult to linearly classify the closer classes. Therefore, non-linear classification hyperplane will help to solve classification problems at that time. SVM is ideally suitable to perform sentiment analysis; because of the sparse nature of text.

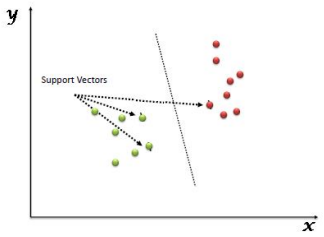

Figure 2:. SVM Hyperplane

In the proposed work, learning model is imported using SVC( ) function of sklearn. Predictions are made on the input test set. After comparing the actual response values with predicted response values, we got the SVM model accuracy as 84.7%.

- Naive Bayes

Naïve Bayes classifiers are a group of probabilistic classifiers, works on Bayes' theorem. It is a family of classifiers which function on a common principle i.e. each feature being classified is independent of the other features. Naive Bayes classifier is a straightforward and powerful algorithm for text classification task such as natural language processing. When we work on a small dataset with some features, Naive Bayes approach is best suitable. It depends on the use of the Baye's standard given by the accompanying equation: $$P(C=c|D=d)=\frac{P(D=d|C=c)P(C=c)}{P(D=d)}$$

Where D denotes the document and C the category (label), d and c are instances of D and

$$P(D=d)=\sum_{c\in C}P(D=d|C=c)P(C=c)$$

We can simplify this expression by,

$$P(c|d)=\frac{P(d|c)P(c)}{P(d)}$$

In our case, a tweet is represented by a vector of attributes such as d = (w1, w2…, wk). Computing P(d|c) is not trivial and that is the reason, the Naïve Bayes introduces the assumption that all of the feature values wj are independent and given the classification label c.

In each machine learning task, it is necessary to have a baseline which is a quick and dirty execution of a fundamental model for doing the principal classification and on the basis of its efficiency, model can be improved. In the proposed work, the Bag of Words model is used to obtain features from the dataset and the classification model obtained an accuracy rate of 80.7%.

- Random Forest

Random Forest is a supervised classification algorithm, which works by creating the forest with a number of decision trees. In general, robustness of the forest depends on the number of trees it has. Similarly, Random Forest technique provides the high accuracy results because of the more number of trees. Number of trees and results obtained

are correlated. Multiple decision trees are built with random forest classifier and then they are merged together to get more exact and stable result. With respect to growing trees, more randomness is added to the model. Here, the best feature is considered among random set of features to develop classification model.

Sklearn provides a great tool to identify the relative importance of individual feature, while performing prediction. Whenever there is a necessity to develop a classification model in a short period of time, random forest technique is more suitable. Here, in the proposed work model provided an accuracy of about 87.2% after execution.

- Logistic Regression

It is a statistical method for data analysis. Here, more than one independent variables are considered to predict the target output. The main objective of logistic regression is, it is a way used to split the data to get an accurate prediction of the class which uses the present information. One of the famous classification algorithms is logistic regression to analyse the target feature. It is a Nonlinear function which uses the sigmoid function as hypothesis which is given by p=1/(1+e-y). Here binary data is taken as the target variable. It is based on binary outcome i.e. 1/0 or Yes/No or True/False. Logistic Regression works well with large dataset.

LogisticRegression() from sklearn.linear_model is used here to import the model. After examining the dataset on the basis of logistic regression, we got the model accuracy as 82.6%.

- K- Nearest Neighbor Classifier

KNN classifier works well on both classification and regression problems. KNN belongs to supervised learning. The algorithm is easily understandable. KNN model is entirely based on the training dataset. Here whenever we require prediction for unseen data, this algorithm will search for k-most similar instances. K-nearest neighbor classifier is powerful because, in order to perform classification it will measure the distance between two instances to find the similarity. Then based on the similarity, it will classify the incoming data.

For implementation of KNN classifier, KneighborsClassifier(n_neighbors=n) function from sklearn.neighbors package is used. Based on the predictions, model provided an accuracy of 75.2%.

3.4.1. Classification Metrics used

Evaluation of classifier performance is carried out with the help of evaluation metrics such as accuracy, precision, recall and F1-score measures. The formulas for measuring these metrics can be stated as:

$$\text{Precision} = \frac{Tp}{Tp+Fp}$$

$$\text{Recall} = \frac{Tp}{Tp+FN}$$

$$\text{Accuracy} = \frac{Tp+TN}{Tp+TN+Fp+FN}$$

$$\text{F1} = \frac{2*Precision*Recall}{Precision*Recall}$$

Where, Tp denotes the total positive tweets which are correctly classified as positive towards Ayodhya issue, TN denotes the number of tweets classified correctly as negative, FN denotes the number of tweets classified wrongly as negative and FP denotes the number of tweets misclassified as positive.

4.2. Confusion Matrix

It is the description of the classification model performance. Confusion Matrix is generated based on test data for which, output label is already known. This gives clear idea about model is giving proper outcome or not. Also it is easy to identify the model errors. We have calculated model accuracy score using confusion matrix.

3.5 Sentiment summation Method

In the sentiment summation method, initially the sentiment score is calculated for individual sentences. Only the words which hold the sentiments are considered for score calculation. Then sentiment aggregation task is performed to get the total sentiment scores. A single sentence may convey positive, negative or neutral opinion towards an object or entity. In the proposed methodology, sentiment score summation is done with respect to Vader Sentiment analyser which distinguishes the sentences into positive or negative based on expressed words. Sentiment score is obtained by calling Polarity_scores() method of Vader Sentiment Analyser. Lexicon ratings are calculated to analyse the sentence score. Sentiment score is between the ranges of -1 to +1.

4. RESULT AND DISCUSSION

The details about the obtained result are discussed here. In the proposed work, we have created a dataset of 15000 tweets, which is distributed in 80:20 ratio between training and testing. Table 1, Table 2, Table 3, Table 4 and Table 5 show the performance measures of Random Forest, KNN, Logistic Regression, Naïve Byes and Support Vector Machine based classifiers respectively. Model efficiencies are compared with respect to precision, recall and F1-score. Similarly, Table 6 shows the performance of the classifier models based on accuracy.

Table 1:. Naive Bayes Classification Measurements

Performance Measures (%)	
Positive Recall	56
Negative Recall	93
Positive Precision	81
Negative Precision	81

Table 2: Random Forest Classification Measurements

Performance Measures (%)	
Positive Recall	76
Negative Recall	93
Positive Precision	84
Negative Precision	89

Table 3: SVM Classification Measurements

Performance Measures (%)	
Positive Recall	66
Negative Recall	94
Positive Precision	85
Negative Precision	85

Table 4:. KNN Classification Measurements

Performance Measures (%)	
Positive Recall	40
Negative Recall	93
Positive Precision	75
Negative Precision	75

Table 5: Logistic Regression Classification Measurements

Performance Measures (%)	
Positive Recall	59
Negative Recall	95
Positive Precision	85
Negative Precision	82

Table 6:. Model Accuracy Comparison

Algorithms	Accuracy (%)
Naïve Bayes	80.7
Random Forest	87.2
Support Vector Machine	84.7
KNN	75.2
LR	82.6

Comparative view of machine learning models is represented in Figure 4.

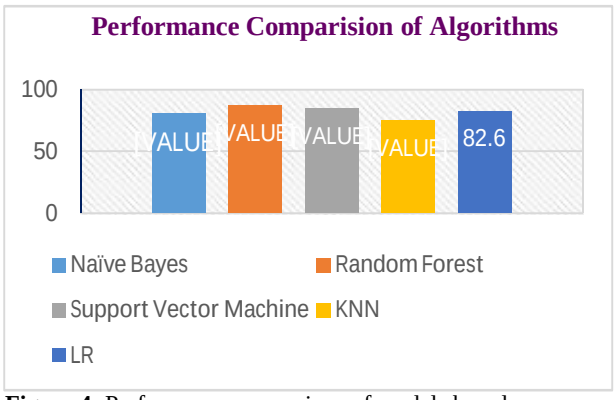

Figure 4: Performance comparison of models based on accuracy

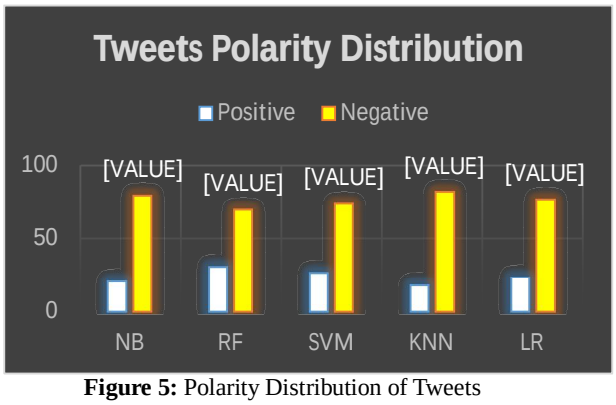

Figure 5: Polarity Distribution of Tweets

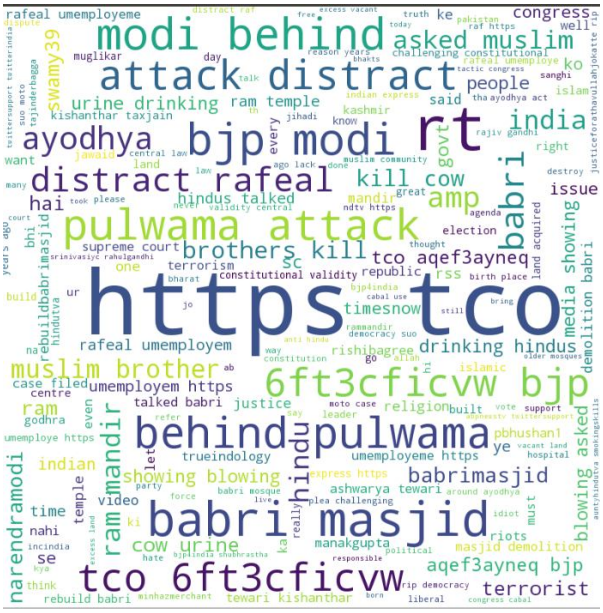

Figure 6: Word Cloud Representation

Figure 6 shows the word cloud representation of data. Words that appear more frequently are given higher importance as per shown representation. It is assumed that the larger visualised word is a more common word in the document. This type of visualization helps in identifying the words which does not express any sentiment values but used commonly. Such identified words can be discounted in the pre-processing step which helps in improving prediction accuracy.

5. CONCLUSION AND FUTURE WORK

Twitter sentiment analysis goes under the classification of text mining. The main focus is to analyse the sentiments of the tweets. Supervised machine learning techniques perform well in sentiment categorisation. In the proposed work, various machine learning models implemented and compared to obtain efficient text categorisation model. From the analysis, it is observed that Random Forest algorithm gives best text classification accuracy as compared to other applied algorithms.

Twitter data is generating tremendously every day. So, there is a difficulty in handling tweets. Therefore, in the future, the plan is to perform twitter sentiment classification on Bigdata framework so that twitter data can be classified within short amount of time.

REFERENCES

[1] Falguni Gupta, Swati Singhal, Amity University, "Sentiment Analysis of the Demonetization of Economy 2016 India, Regionwise", 2017 7th International Conference on Cloud Computing, Data Science & Engineering, January 2017, pg. 693-696.

[2] Geetika Gautam, Divakar Yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and

Semantic Analysis", 2014 Seventh International Conference on Contemporary Computing (IC3), August 2014, pg.437-442.
[3] Nicolas Tsapatsoulis, Constantinos Djouvas, "Feature Extraction for Tweet Classification: "Do the Humans Perform Better?" 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), July 2017, pg. 53-58.
https://doi.org/10.1109/SMAP.2017.8022667
[4] Taiwo Kolajo, Olawande Daramola, Ayodele Adebiyi, "Sentiment Analysis on Naija-Tweets", in the proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.
https://doi.org/10.18653/v1/P19-2047
[5] Huma Parveen, Prof. Shikha Pandey, "Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorithm", 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), July 2016, pg. 416-419.
[6] Omar Abdelwahab, Mohamed Bahgat, Christopher J. Lowrance, Adel Elmaghraby, "Effect of Training Set Size on SVM and Naïve Bayes for Twitter Sentiment Analysis", 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), December 2015, pg.46-51.
[7] Neethu M S, Rajashri R, "Sentiment Analysis in Twitter using Machine Learning Techniques", 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), July 2013,pg. 1-5.
[8] Tapan Sahni, Chinmay Chandak, Naveen Reddy Chedeti, Manish Singh, "Efficient Twitter Sentiment Classification using Subjective Distant Supervision", 2017, 9th International Conference on Communication Systems and Networks (COMSNETS), January 2017, pg. 548-553.
[9] Shohreh Haddadan, Elena Cabrio and Serena Villata, "Mining Arguments in 50 Years of US Presidential Campaign Debates", in the proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.
https://doi.org/10.18653/v1/P19-1463
[10] Anjli Bhavan, Rohan Mishra, Pradyumna Prakhar Sinha, Ramit Sawhney, Rajiv Ratn Shah, "", in the proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.