IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol-6, Special Issue-6, July 2018 E-ISSN: 2347-2693

Opinion Mining on Twitter Data Using Supervised Machine Learning Algorithms

Deepa Mary Mathews[1*], Sajimon Abraham[2]

[1]Research Scholar, Mahatma Gandhi University, Kottayam, India
[2] School of Management Studies, Mahatma Gandhi University, Kottayam, India

Available online at: www.ijcseonline.org

***Abstract*—** The emerging digital era generates heaps of computerized information. The greater part of the electronic data in the world today has been created over the last recent couple of years. The velocity of data generation is unimaginable and incomprehensible. People nowadays are commonly using the digital media to express their stand point about a topic. These opinions are analyzed automatically to know whether the client remark is ideal or not good to the said theme. This ought to be possible by Opinion Mining, also called as Sentiment Analysis. The basic chore in Sentiment Analysis is to categorize the orientation of a given review and subsequently identifying whether the sentiment implied is positive, negative or fair. In this paper, the tweets based on the news thread "Whether National Anthem is needed at Cinema theatres?" are analyzed based on the user rating for the opinions. The classifiers like Bernoulli and Multinomial Naive Bayes, Random Forest, k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) have been used for analyzing the opinions and found that the Random Forest classifier and Multinomial Naïve Bayes classifier is the top rated classifier based on their accuracy values.

***Keywords*—** Sentiment Analysis, Naive Bayes Classifier, SVM, Random Forest Classifier, KNN Classifier

I. INTRODUCTION

We exist in an advanced time where information is intensifying quickly as a result of the increasing utilization of the web, sensors, and demonetization and so on. As per the statistics based on internetlivestas.com, in a second there are 8039 tweets, 1,377 Tumblr posts, 136,000 photographs transferred on Facebook, 56,844GB of Internet traffic, 2,691,713 Emails sent etc shows the significance of technologies and information systems used in this era. Major part of these data is in the textual format and is unstructured. With the tremendous accessibility of archives which express suppositions on various issues, the challenge arises to analyze it and produce useful knowledge detach from it. The exploration of these users created content and the exact aspect of client standpoints towards items and events is rather valuable to many applications. To break down this large amount of data, different opinion mining techniques can be used.

The Opinion mining leads to dig out the user perspectives, outlook, feeling and sentiments from the user generated data. The process identifies the direction of a given text data to know whether the communicated sentiment is sure, negative, or fair. The authors here evaluated the sentiments of the users who posted their opinions about the Supreme Court decision – "Whether National Anthem is to be played in the Cinema theatres?". The reviews are taken from the Twitter. The authors explored the possible ways to analyze the user sentiments using Python programming libraries and various Machine Learning algorithms. The classifiers like Bernoulli and Multinomial Naive Bayes, Random Forest, k-NN and SVM have been used for analyzing the opinions. The authors performed a comparison among these classifiers and found that the Random Forest classifier and Multinomial Naive Bayes classifier is the top rated classifier based on their accuracy values on validation set.

Rest of the paper is organized as follows, Section II contain the literature survey done related to the topic, Section III explain the methodology and the algorithm used, Section IV describes results and discussion, Section V concludes research work with future directions.

II. RELATED WORK

All through recent years, a recurrent number of discussions learning varied class of opinion mining in English annals have been observed. Cases of such sorts incorporate objectivity and subjectivity recognition, polarity classification, perspective based conclusion arrangement and so forth. Various methodologies used for opinion mining can be found in [5, 6, 7]. In the article [2], authors concluded that classification models can be selected based on resources, accuracy requirement, training time available etc. Ankita Gupta et.al introduced a hybrid model of k-Nearest Neighbor and Support Vector Machine in [3]. Lopamudra Dey [4], et.al, in their paper performed Sentiment Analysis using SVM and kNN classifier. In [1], the authors proposed a method that used machine learning methods based on opinion mining to get opinions from the text documents.

III. Methodology

The methodologies used for learning and analyzing the sentiments are elucidated in this section. We portrayed different methodologies used in the study to retrieve the tweets, to pre-process the sentiments and the various classification algorithms used to analyze the sentiments.

A. Retrieval of Tweets using Twitter API

Twitter is one among the mostly used microblogging website with about 33 billion active users. As a result of its wide prominence, the velocity of generation of tweets is around 340 million every day. Tweets can be seen as the good response of what is occurring far and wide. It allows the dissemination of user opinion about the latest trend and thread. Twitter developer API provides facility to get to and recover twitter information. To get access to the Twitter API and thereby the tweets, the credentials like application consumer key, consumer secret key, access token key and the access token secret key is required. Using Python programming language which uses these API credentials, the tweets are acquired.

The percentage of positive, negative and neutral reviews is shown in figure 1. The statistics of the dataset is shown in Table 1. The dataset contains the tweets based on the news thread "Whether National Anthem is to be played at Cinema theatres?" extracted during March, 2018. The number of features extracted is 1055. The dataset is trained based on the user ratings.

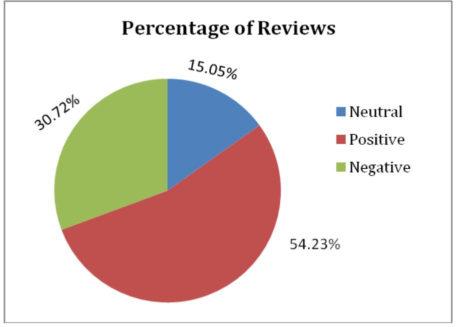

Figure 1. Percentage of Reviews in the Dataset

Table 1: Statistics of the Dataset

Number of Reviews	319
Number of features	1055
Percentage of Positive Reviews	54.23%
Percentage of Negative Reviews	30.72%
Percentage of Neutral Reviews	15.05%

B. Preprocessing the tweets using NLTK packages

Social media reviews don't tag on any grammar, mainly short messages, spell faults are common, and have many irrelevant words. So a pre-processing phase is required. It is the way toward acquainting another document to the system, where every standpoint of the client corresponds to a set of key terms for the efficient storage and retrieval of the data. The data from reviews are tokenized into convenient units to build a representation of the data. This phase removes the data which is not required for analysis. NLTK package is used for pre-processing texts.

NLTK (Natural Language Tool Kit), a platform for putting up Python programs to cope with human language, provides various packages for pre-processing the reviews [5]. Various pre-processing steps done to clean/ correct tweets retrieved includes

- Tokenization
- Removal of special characters, Punctuation marks
- Removal of numbers
- Removal of website links
- Striping off whitespaces
- Removal of Stopwords
- Lemmatization and Stemming
- Drop out the missing values, if any

Tokenization is the process of slicing up the sentences [9]. The additional spaces, special characters, digits, hash tags, URLs and so forth are removed. Words are lemmatized to the fundamental form using the lemmatize method and stemming is done using Snowball stemmer.

C. Classification Algorithms

Classification algorithms that use supervised approach, like Multinomial Naive Bayes (MNB), SVM, Bernoulli Naive Bayes (BNB), Random Forest (RF) and k-NN are used in this article to perform the Sentiment analysis on the review dataset.

Naive Bayes classifiers deduce that the classes for classification are independent [2]. This classifier is usually used when the time taking for labeling is a crucial aspect. The Multinomial Naive Bayes classifier is applicable to the situations when several occurrences of the term signify in the classification. The Bernoulli Naïve Bayes classifier is mostly employed when the nonexistence of a term matters.

The Support Vector Machine is based on ruling a parting between hyper planes distinct by the classes of data [10]. It has the ability to learn independent of the dimensionality of the feature space. The Linear SVC is used here.

k-Nearest Neighbors algorithm is a classification algorithm which considers the likeness between the k adjoining neighbors [2]. In this algorithm, an item is classified by mass choice of its neighbors. The value of k is taken as two in this article.

Random forests classifier brings out multi-altitude decision trees. The relation amid the trees is shortened by arbitrarily choosing trees [11] and accordingly the prediction exactness raises and thereby increases the effectiveness.

IV. IMPLEMENTATION

The user review document is pre-processed and the missed values are removed, if any. Based on the user ratings, the reviews are classified. The user rating 3 is considered as neutral and is removed. The ratings 4s and 5s are encoded as positive and 1s and 2s are encoded as negative. The 10% portion of this labelled dataset is used for validation and the rest as training set. The document term matrix is generated. The feature set generated is then trained using classification algorithms and the validation set is validated using these classifiers.

The framework of implementation can be shown as

i) Load the Pre-processed dataset
ii) Remove the ‘Neutral’ reviews (ratings equal to 3)
iii) Encode the ratings 4s and 5s as 1 (Positive Sentiment) and 1s and 2s as 0 (Negative Sentiment)
iv) Split the dataset into training set and test set
v) Fit and transform the training data to a document-term matrix
vi) Train Multinomial Naive Bayes, Bernoulli Naive Bayes, Support Vector Machine, Random Forest and k-Nearest Neighbour Classifier
vii) Evaluate the classifiers on the test set and find out the Accuracy and AUC score
viii) Plot the value of evaluation measures of the 5 classifiers

V. EVALUATION MEASURES

Classifiers are learned or trained on a finite training set. A learned classifier has to be tested on a different test set experimentally. The experiments on test set are a deputy for the performance on invisible data which ensure the classifier’s generalization ability. The criterion function used for evaluating the classifier performance experimentally, are accuracy, AUC score etc. This classification produces four outcomes - the correct and incorrect positive and negative prediction called as TruePositive, TrueNegative, FalsePositive and FalseNegative. Based on these four values, a confusion matrix can be created.

Various evaluation measures derived from the Confusion Matrix are

i) Precision is the figure of accurate positive predictions among the total positive predictions

$$Precision = TruePositive/ (TruePositive + FalsePositive)$$

ii) Recall is part of positive (P) class samples accurately classified.

$$Recall = TruePositive/P$$

iii) F-Score is the harmonic mean of precision and recall.

$$F-Score = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

Table 2. Precision, Recall, F-Score values

Classifier	Precision	Recall	F-Score
Bernoulli NB	0.77	0.79	0.75
SVM	0.93	0.93	0.93
Random Forest	0.97	0.96	0.96
kNN	0.83	0.82	0.83
Multinomial NB	0.97	0.96	0.96

The Precision, Recall, F Score values generated using the five classification algorithms is shown in the Table 2 and its graphical representation in figure 2.

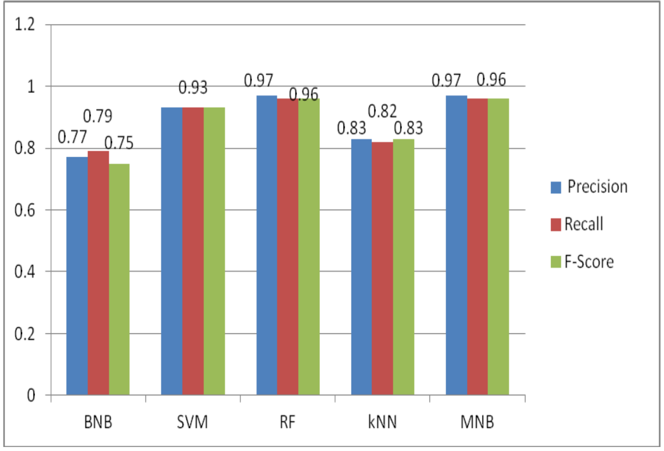

Figure 2. Precision, Recall and F-Score values of five classifiers

VI. RESULTS AND DISCUSSION

The accuracy of the classifier is calculated as the total number of two correct predictions (TP + TN) divided by the total number of a dataset (P + N). The AUC means Area under Curve. It is used in classification analysis in order to determine which of the used models predicts the classes best. An example of its application is ROC curves. Here, the true positive rates are plotted against false positive rates.

The Accuracy value and the AUC Score calculated using Bernoulli Naive Bayes, SVM, Random Forest, k Nearest Neighbor and Multinomial Naïve Bayes classification algorithms done on the Validation set are shown in the figure 3 and in Table 3. The values shows that the accuracy and the AUC score are higher if Random Forest Classifier is using.

Table 3. Accuracy and AUC Score values

Classifier	Accuracy	AUC score
Bernoulli NB	78.57%	61.90%
SVM	92.86%	90.48%
Random Forest	96.43%	92.86%
kNN	82.14%	78.57%
Multinomial NB	96.43%	92.86%

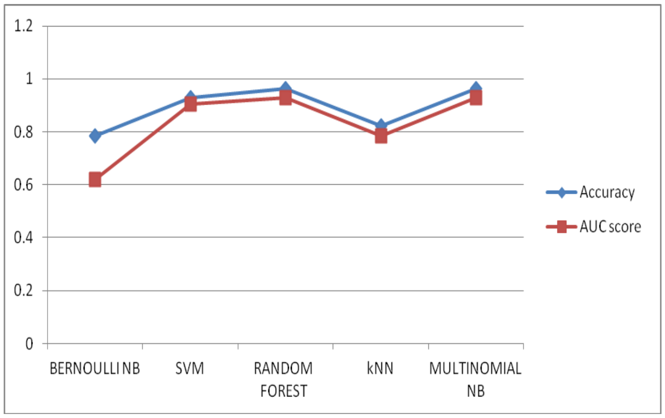

Figure 3. Accuracy and AUC values

VII. Conclusion and Future Scope

The experimentation result shows that if accuracy is the main concern then we ought to choose a classifier like Random Forest or Multinomial Naive Bayes classifier. The authors used various Supervised Machine Learning algorithms for Sentiment Analysis and are found that most of the users are supporting the decision taken by Supreme Court regarding "whether National Anthem is to be played in the Cinema theatres". The major limitation is that the training set may contain misspelled or mislabelled values which may mislead and affect the performance of the classifier.

Unsupervised algorithms, which are not based on the training or labelled set can also, be used for the Sentiment Analysis. The authors considered only one constraint here for choosing the classifier. The choice of classifiers will also be determined based on resources, time required for learning and training and so forth, which is a future scope.

References

[1] Khan, Khairullah, Baharum B. Baharudin, and Aurangzeb Khan. "Mining opinion targets from text documents: A review." Journal of Emerging Technologies in Web Intelligence 5.4 (2013): 343-353.

[2] Gupte, Amit, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, and A. Gupte, "Comparative study of classification algorithms used in sentiment analysis.", *International Journal of Computer Science and Information Technologies* , Vol. 5 (5) : 6261-6264, 2014

[3] Gupta, Ankita, Jyotika Pruthi, and Neha Sahu. "Sentiment Analysis of Tweets using Machine Learning Approach." (2017).

[4] Dey, Lopamudra, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, and Sweta Tiwari. "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier." *arXiv preprint arXiv:1610.09982* (2016).

[5] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5, no. 1 (2012): 1-167.

[6] Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis." In *Mining text data*, pp. 415-463. Springer US, 2012.

[7] Mohammad, Saif M. "Sentiment analysis: Detecting valence, emotions, and other affectual states from text." In *Emotion measurement*, pp. 201-237. 2016.

[8] Wagner, Wiebke. "Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit." *Language Resources and Evaluation* 44, no. 4 (2010): 421-424.

[9] Palmer, David D. "Tokenisation and sentence segmentation." *Handbook of natural language processing* (2000): 11-35.

[10] Ye, Qiang, Ziqiong Zhang, and Rob Law. "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches." Expert systems with applications 36.3 (2009): 6527-6535.

[11] Aldoğan, Deniz, and Yusuf Yaslan. "A comparison study on active learning integrated ensemble approaches in sentiment analysis." Computers & Electrical Engineering 57 (2017): 311-323.

Authors Profile

Ms. Deepa Mary Mathews, is presently Assistant Professor of the Department of Computer Applications, FISAT. She was graduated in Chemistry from Mahatma Gandhi University, had her post graduation in Computer Applications from Madurai Kamaraj University, and gained second post graduation in M.Tech in Computer Science and Engineering from Dr.M.G.R.University in the year 2006. She is currently pursuing Ph.D and her research area includes Data Mining, Social Data Analytics and Machine Learning. She has published 9 research papers in the International Journals, National and International Conferences.

Mr. Sajimon Abraham, MCA, MSc(Mathematics), MBA, PhD (Computer Science). He has been working as Faculty Member in Computer Applications & IT, School of Management and Business Studies, Mahatma Gandhi University, Kottayam, Kerala, India. He currently holds the additional charge of Director(Hon), University Center for International Co-operation. He was previously working as Systems Analyst in Institute of Human Resource Development, Faculty member of Computer Applications in Marian College, Kuttikkanam and Database Architect in Royal University of Bhutan under Colombo Plan on deputation through Ministry of External Affairs, Govt of India. His research area includes Data Science, Spatio-Temporal Databases, Mobility Mining, Sentiment Analysis, Big Data Analytics and E-learning and has published 52 articles in National, International Journals and Conference Proceedings.