International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-8, Issue-11, 2021 Regular Edition E-ISSN: 2349-7084

Twitter Sentiment Analysis Using Machine Learning

[1]Pasunooru Santosh Reddy, [2]Bheempaka Sai Kumar, [3]Reddy Reddy Jahnavi Reddy

[1,2] IV Year,CSE Dept,CVR College of Engineering, Vastunagar, Mangalpalli (V), Ibrahimpatnam (M), Rangareddy (D), Telangana , India
[3]SASTRA (DEEMED TO BE UNIVERSITY), Thanjavur, Tamil Nadu, India
.
Email ID: pasunoorusantoshreddy@gmail.com, sai91457@gmail.com, reddyjanur@gmail.com
**Corresponding Author: pasunoorusantoshreddy@gmail.com*

Available online at: http://www.ijcert.org

Received: 08/11/2021, Revised: 12/11/2021, Accepted: 16/11/2021, Published: 18/11/2021

Abstract:- Sentiment analysis is the process of identifying and categorising the emotions expressed in text. When tweets are analysed, they typically generate a large amount of sentiment data. This information allows us to better understand people's perspectives on a variety of issues. This study tries to classify tweets based on their sentiment. They can express either positive or negative emotions. Twitter is a social networking and micro blogging platform that allows users to post 140-character status updates or opinions. It has about 200 million registered users, 100 million active users, and half of them log in every day, resulting in nearly 250 million tweets per day. Because of the widespread use, we want to reflect the prevalent attitude by analysing tweets. Predicting political elections and macroeconomic phenomena like stock exchanges necessitates a look at public sentiment. We attempt to categorise the tweets as positive or negative. To represent the "tweet," it must also extract valuable elements from the text, such as unigrams and bigrams. We use machine learning methods to analyse sentiment using the collected features. Individual models did not provide high accuracy on their own. So we created an Ensemble Model that predicts based on a majority vote using Naive Bayes, Logistic Regression, and Support Vector Machines.

Keywords: Sentiment analysis, twitter, Machine Learning , Ensemble Model, Naive Bayes, Logistic Regression, and Support Vector Machines.

---

1. Introduction

Twitter Sentiment Analysis was thoroughly dealt by Alec Go, Richa Bhayani and Lei Huang, Computer Science graduate students of Stanford University. They used various classifers, including Naive Bayes, Maximum Entropy as well as Support Vector Machines to classify the Tweets. The feature extractors used by them were both unigrams and bigrams combined. Parts of speech tag was used because same word may have different meaning depending on its usage. The data-set used by them was huge, comprising 1.6 million tweets divided equally into positive and negative classes. We have also used the same dataset, but considered only 200,000 tweets with 100,000 positive and 100,000 negative tweets. This was done to cut down the running time of the program. Also, unigrams and Bigrams have been adopted as feature extractors to reduce the size of the feature vector. The classifiers used by us to create an Ensemble model are Naive Bayes, Logistic Regression and Support Vector Machines.

Sentiment Analysis: Sentiment Analysis is the interpretation and classification of emotions within text data using text analysis techniques. Most of the research in sentiment analysis has been aimed at larger pieces of text, like movie reviews, or product reviews Consumers can use sentiment analysis to research products and services before a purchase. Production companies can use the public opinion to determine acceptance of their products and the public

 https://doi.org/10.22362/ijcert/2021/v8/i11/v8i1101

demand. Movie-goers can decide whether to watch a movie or not after going through other people's reviews.

Twitter Sentiment Analysis: Traditionally, most of the research in sentiment analysis has been aimed at larger pieces of text, Like movie reviews, or product reviews. Tweets are more casual and are limited by 140 Characters. However, this alone does not make it an easy task (in terms of programming time, not in accuracy as larger piece of text tends to be correctly classified) as people rarely give a second thought before posting a tweet. Grammar and content both suffer at the hands of the tweeter. The presence of a large dataset is always recommended (for better training of the classifier) and twitter makes it possible to obtain any number of tweets during a desired period. However, various difficulties are faced during processing of raw tweets.

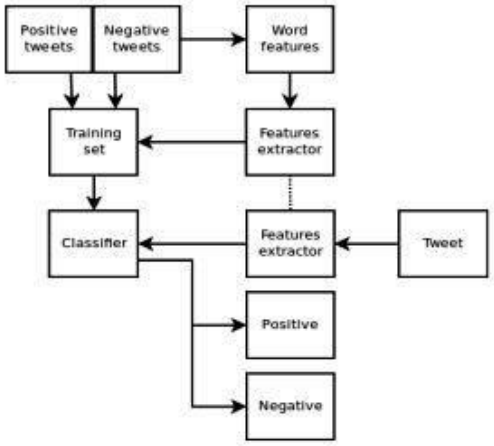


Figure 1. General model for sentiment analysis

Problem Statement: Given a message, classify whether the message is of positive or negative sentiment. For Messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen. A problem In sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level. Whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, neutral. Despite the availability of software to extract data regarding a person's sentiment on a specific product or service, organizations and other data workers still face issues regarding the data extraction. Despite the availability of software to extract data regarding a person's sentiment on a specific product or service, organizations and other data workers still face issues regarding the data extraction. Despite the availability of software to extract data
regarding a person's sentiment on a specific product or service, organizations and other data workers still face issues
Regarding the data extraction.

Despite the availability of software to extract data regarding a person's sentiment on a specific Product or service, organizations and other data workers still face issues regarding the data extraction. Sentiment Analysis of Web Based Applications Focus on Single Tweet Only. With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis. This translates to a huge volume of information from a human viewpoint
Which make it difficult to extract sentences, read them, analyse tweet by tweet, summarize?
Them and organize them into an understandable format in a timely manner.
The main contribution of the paper is as follows

- We train and test models individually to know the accuracy By performing Regularization and Hyper parameter tuning, we try to improve accuracy of models.
- For hyper parameter tuning we are using gridSearchCV and finding best suitable parameters for each algorithm.
- To achieve high Accuracy, we are creating an ensemble model using different machine learning classifiers i.e Logistic Regression, Naive Bayes, and Support Vector Machines (SVM). It predicts based on majority voting of individual models.

2. Related Work

According to the author, an ensembles classification has been used to upgrade the correctness of tweet sentimental classification. Companies detect the consumer's interest to choose products and brands based on their opinions. It is a type of blog where the user can create, post, and update as well as read the short messages. [1]

Deep convolution algorithm is used for better performance in twitter analysis. The survey is taken on public reviews about the related product and events. Word embedding method institute individual learning based on twitter. Twitter sentiment is well performed using pre-trained word vector. [2] [3].

Deals with Tweets to polls where sentiments are measured from text based on public opinion. Advanced NLP techniques are useful to enhance opinion estimation. Textual sentiment in tweet message is measured through time, comparing to contemporaneous polling data. Author proposed Naive Bayes classification algorithm for both sarcastic and non-sarcastic tweets is the form of expressing negative feeling using positive words. In the existing system, logistic regression technique is used to detect sarcasm in tweets, it has some drawbacks which cannot predict continuous variables [4] [5].

In this paper collection of written texts and dictionary based methods are used to determine both positive

and negative words in tweets. The work presented in paper specifies new techniques for sentiment analysis on twitter data. The overall tweet sentiment was calculated using a linear equation. Several steps were taken for sentimental analysis on twitter data using machine learning algorithm. Tweets are pre-processed using NLP based techniques. They build the model using Support Vector Machine (SVM), Naive Bayes classifiers, and machine learning classifiers. The outcome shows that decision tree performs effectively showing 100% accuracy. [6] Moreover, substantial work put forward a solution for sentimental analysis based.

The researchers used data consisting of tweets with emoji's. Finally, model was trained using Maximum Entropy, Naive Bayes classifiers, Support Vector Machine (SVM). The outcome shows that SVM model was more effective than other models [7].

Data analysis of twitter includes machine learning and lexicon-based approach. Comparatively, various research being done with sentiment. The study defines the concept of opinion in sentiment analysis in twitter. The result shows that machine learning method such as Naive Bayes and SVM have the highest accuracy, it improves the robustness and performance of twitter. [8].

Sentimental analysis is used for classifying the positive and negative opinion using Machine Learning Techniques with the help of SVM algorithm which shows maximum accuracy. Sentimental analysis is most effectively used for analyzing the people opinion. It is used for result comparison purpose. [9] .

Tweet sentimental analysis has been undiscovered in the literature. Marketing the product which is useful for the customer to search the products and brand. Here, Automatic tool is used for classifier ensembles and lexicons. This analysis is used for classifying problems like opinions, attitudes, and emotions. [10]

3. System Study

Existing System: Various researchers have been working on twitter and from time to time they are publishing their researches. They have used various sentiment analysis techniques for improving the results of classification their work is also helpful in this research as the sentiment analysis techniques they have used, feature selection techniques, different pre-processing steps they have used is taken care of in this research [12]. This research mainly focuses on supervised approach for sentiment analysis task and has surveyed researches both for twitter and non- twitter data and also for both supervised and lexicon based approaches for better clarification and understanding of the topic chosen. Many researches defined multiple faces of sentiment analysis as opinion orientation, feature extraction etc[13-15]. Machine learning classifiers need various features for learning so different researchers from time to time have selected different features for comparing results. Different features and feature selection methods as semantic features and concepts, information gain, chi-square etc[16].

Challenges with the existing works : While computer systems and machine learning algorithms are getting better all the time, they still face challenges when deciphering human sentiments in online statements.

Proposed System:

Proposed architecture is shown in figure 2.

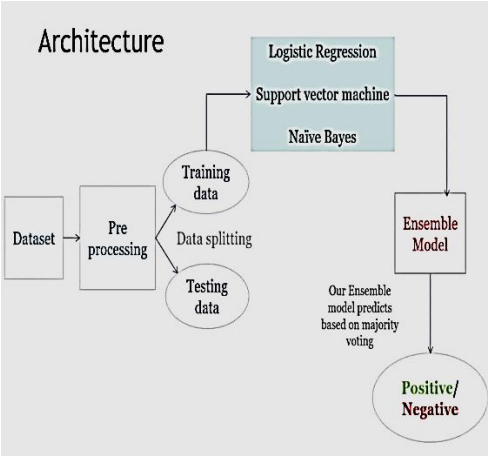


Figure 2. Proposed architecture

The system shows that initially dataset will consider for the experiment, where data pre-processing is required to remove the duplicate of records and address the missing and mull values. After completion of pre-processing dataset will be divided as training data and test data after that model will be created and apply logistic Regression , Support Vector Machine and Naïve Bayes algorithm classification model , finally data will be classified as positive and negative values.

4. Results and Analysis

Dataset: This is the sentiment140 dataset. It contains 22500 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment .It contains the following 6 fields:

target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)

ids: The id of the tweet (2087)

date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)

flag: The query (lyx). If there is no query, then this value is NO_QUERY.

user: the user that tweeted (robotickilldozr)

text: the text of the tweet (Lyx is cool)

Logistic Regression:

Accuracy of model on Training Data: 87.92 %

Accuracy of model on Testing Data : 77.73%

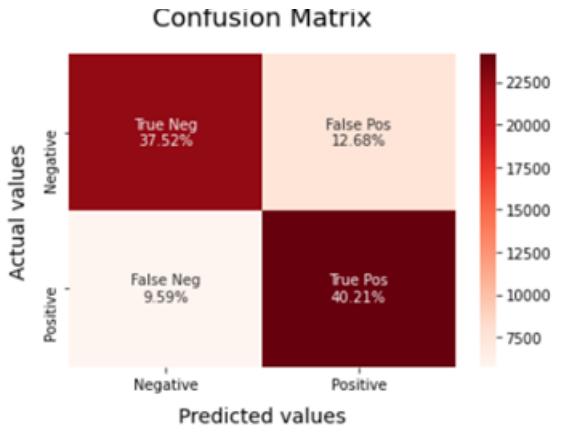
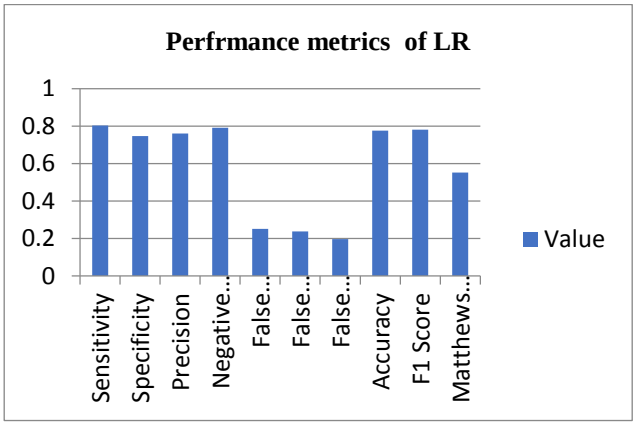


Figure 3. Confusion Matrix of Logistic Regression

Table 1. Performance metrics of Logistic Regression

Measure	Value	Derivations
Sensitivity	0.8074	TPR = TP / (TP + FN)
Specificity	0.7474	SPC = TN / (FP + TN)
Precision	0.7603	PPV = TP / (TP + FP)
Negative Predictive Value	0.7964	NPV = TN / (TN + FN)
False Positive Rate	0.2526	FPR = FP / (FP + TN)
False Discovery Rate	0.2397	FDR = FP / (FP + TP)
False Negative Rate	0.1926	FNR = FN / (FN + TP)
Accuracy	0.7773	ACC = (TP + TN) / (P + N)
F1 Score	0.7831	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.5558	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))



Support Vector Machine:

Accuracy of model on Training Data : 88.02 %

Accuracy of model on Testing Data : 77.80 %

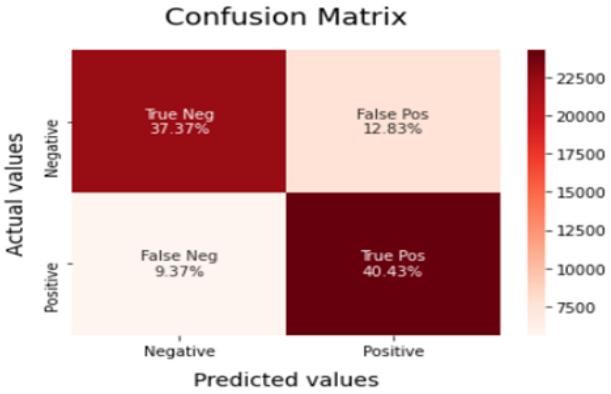
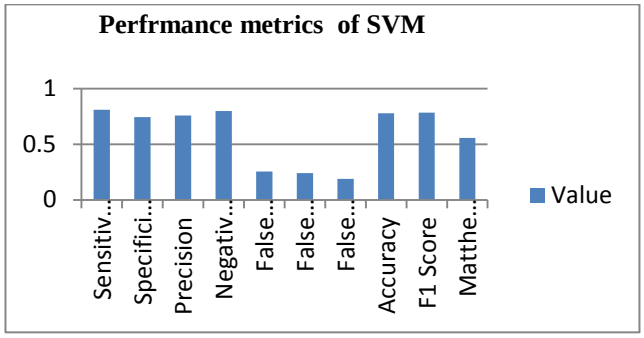


Figure 4. Confusion Matrix of Support Vector Machine

Table 2. Performance metrics of Support Vector Machine

Measure	Value	Derivations
Sensitivity	0.8119	TPR = TP / (TP + FN)
Specificity	0.7445	SPC = TN / (FP + TN)
Precision	0.7591	PPV = TP / (TP + FP)
Negative Predictive Value	0.7995	NPV = TN / (TN + FN)
False Positive Rate	0.2555	FPR = FP / (FP + TN)
False Discovery Rate	0.2409	FDR = FP / (FP + TP)
False Negative Rate	0.1881	FNR = FN / (FN + TP)
Accuracy	0.7780	ACC = (TP + TN) / (P + N)

F1 Score	0.7846	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.5575	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))



Bernoulli (Naïve Bayes):

Accuracy of model on Training Data : 94.59 %

Accuracy of model on Testing Data : 76.21 %

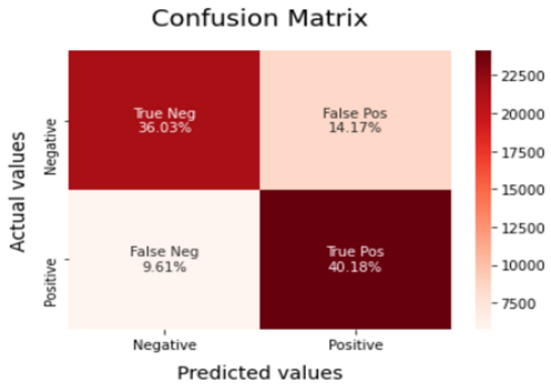
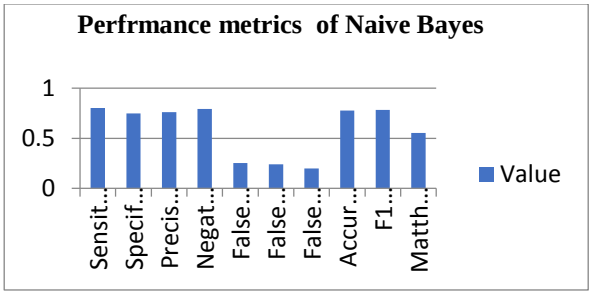


Figure 5. Confusion Matrix of Bernoulli (Naïve Bayes)

Table 3. Performance metrics of Bernoulli (Naïve Bayes)

Measure	Value	Derivations
Sensitivity	0.8070	TPR = TP / (TP + FN)
Specificity	0.7180	SPC = TN / (FP + TN)
Precision	0.7393	PPV = TP / (TP + FP)
Negative Predictive Value	0.7897	NPV = TN / (TN + FN)
False Positive Rate	0.2820	FPR = FP / (FP + TN)
False Discovery Rate	0.2607	FDR = FP / (FP + TP)
False Negative Rate	0.1930	FNR = FN / (FN + TP)
Accuracy	0.7623	ACC = (TP + TN) / (P + N)
F1 Score	0.7717	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.5270	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))



Ensemble Model:

Accuracy of model on Training Data : 88.24 %

Accuracy of model on Testing Data : 77.6 %

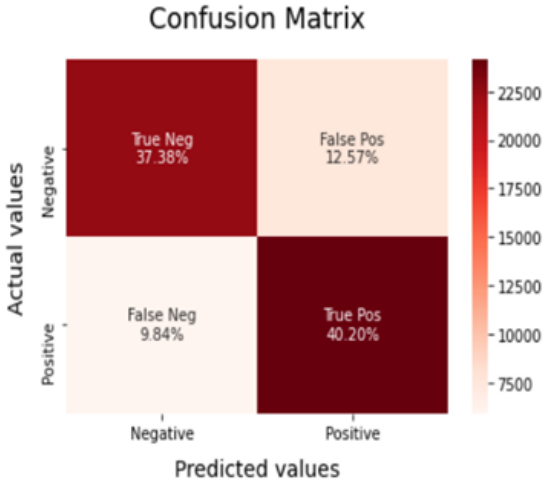
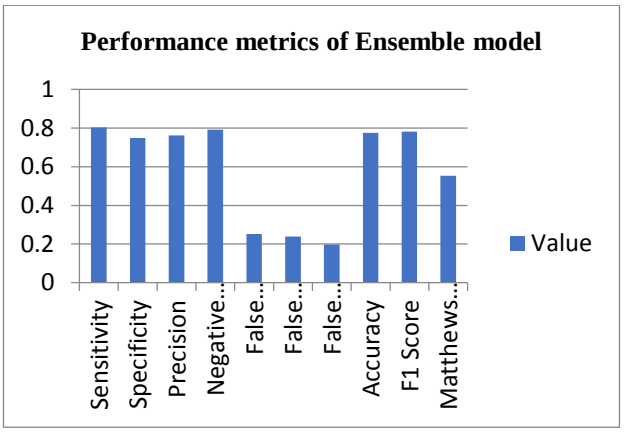


Figure 6. Confusion Matrix of Ensemble Model:

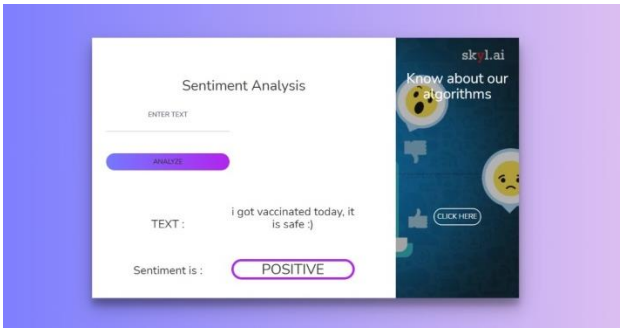
Table 4. Performance metrics of Ensemble Model

Measure	Value	Derivations
Sensitivity	0.8034	TPR = TP / (TP + FN)
Specificity	0.7484	SPC = TN / (FP + TN)
Precision	0.7618	PPV = TP / (TP + FP)

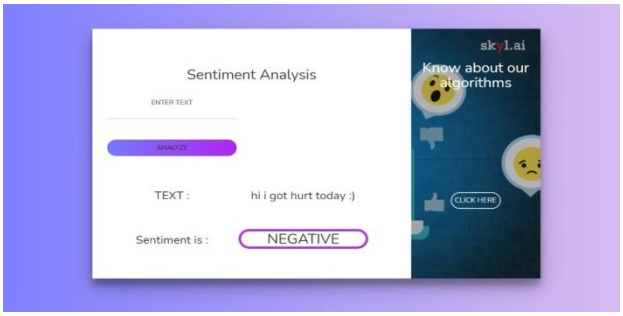
Negative Predictive Value	0.7916	NPV = TN / (TN + FN)
False Positive Rate	0.2516	FPR = FP / (FP + TN)
False Discovery Rate	0.2382	FDR = FP / (FP + TP)
False Negative Rate	0.1966	FNR = FN / (FN + TP)
Accuracy	0.7759	ACC = (TP + TN) / (P + N)
F1 Score	0.7820	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.5526	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))



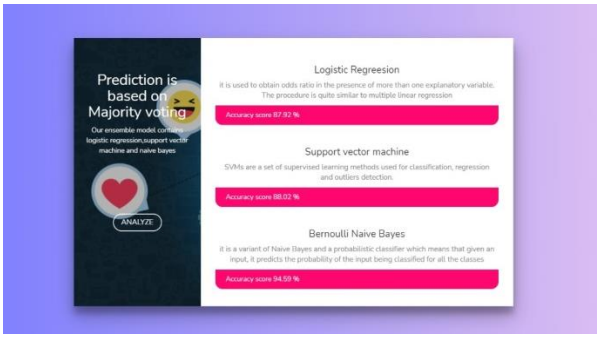
Response for positive tweet



Response for Negative tweet



About algorithms page



5. Conclusion and Future Scope

In tweets data replacing emoticons with their emotion for training data is an effective way to improve the vocabulary of models. Machine learning algorithms (Naive Bayes, Logistic Regression, and support vector machines) can achieve high accuracy for classifying sentiment when using this method. Machine learning techniques perform well for classifying sentiment in tweets. We believe that the accuracy could still be improved. The polarity of a tweet may depend on the perspective you are interpreting the tweet from. Our best classifier has an accuracy of 94.3% for tweets across all domains. If limited to particular domains (such as movies) we feel our classifiers may perform better. Handling neutral tweets In real world applications, neutral tweets cannot be ignored. Proper attention needs to be paid to neutral sentiment. Internationalization We focus only on English sentences, but Twitter has many international users. It should be possible to use our approach to classify sentiment in other languages.

In Future, We can add two more machine learning algorithms to our Ensemble model like KNN algorithm, Decision trees etc. Also we can use Soft voting technique to our ensemble model which predicts based on probabilities for class labels and predicts the class label with the largest sum probability.

References

[1] Ankita, Nabizath Saleena "An ensemble classification system for twitter sentimental analysis",International Conference On Computational System for Twitter Sentiment Analysis (ICCIDS) Procedia Computer Science 132(2018)937-946

[2] Zhao Jianqiang , Gui Xlaolin, and Zhang Xuejun "Deep convolution neutral network for twitter sentimental analysis" Jan 1,2018 ACCESS.2017.2776

[3] Brendan o' Connor , Ramnath Balasubramaniyan , Bryan R. Routledge , Noah A. Smith " Linking Text Sentimental to public opinion time series" International Conference On Web and Social Media(ICWSM),2010,volume 11,nos.122-129,pp.1-2

[4] Bala Durga Dharmavarapu,Jayang Bayana "Sarcasm dection in twitter using sentimental analysis" International Journal of Recent Technology and Engineering(IJRTE) volume 8,issue -1 may 2019

[5] Akshi Kumar and Teeja Mary Sebastian "sentiment analysis on twitter" International Journal of computer

[6] P.Nakov,A.Ritter,S.Rosenthal, F.Sebastiani, and V.Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," in proc.10th Int.Work.Semant.Eval.,Jun.2016,pp1-18.

[7] I.H.Witten, E.Frank,M.A.Hall, and C.J.pal, Data Mining: PracticalMachine Learnimg Tools and Techniques.San Mateo,CA.USA:Morgn Kaufmann,2016

[8] Avinash Surnar, Sunil Sonawane "Review for twitter sentiment analysis using various method" International Journal Of Advanced Research In Computer Engineering And Technology(IJARCET) Volume 6,Issue 5, May 2017

[9] P. A. Gutierrz, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez, "Ordinal regression methods: Survey and experimental study," IEEE Trans. Knowl. Data Eng., vol.28, no. 1,pp. 127-146, jan. 2016.

[10] N. M. Dhanya and U.C Harish "Sentimental analysis on twitter data on demonetization using machine learning techniques" Springer International Publishing AG 2018 DOI:10.1007/978-3-319-71767-8_19

[11] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. In CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, New York, NY, USA, 2009. ACM

[12] G. Mishne. Experiments with mood classification in blog posts. In 1st Workshop on Stylistic Analysis Of Text For Information Access, 2005.

[13] K. Nigam, J. Larerty, and A. Mccallum. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999.

[14] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008.

[15] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.

[16] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009)

