Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Text based Tweet Classification using Ensemble Classifier

Ismankhan Y M

PG Scholar, Department of Computer Science and Engineering, Government College of Technology, Coimbatore, India

Email: isma.71772177102@gct.ac.in

Abstract

There are so many social networking sites available. Tweets have evolved into a crucial tool for gathering people's thoughts, ideas, behaviours and sentiments surrounding particular entities. One of the most intriguing subjects in this context is analyzing the sentiment of tweets using natural language processing (NLP). Although several methods have been created, the accuracy and effectiveness of those methods for sentiment analysis are yet to be improved. This paper proposes an innovative strategy that takes advantage of machine learning and lexical dictionaries. Tweets are classified using a stacked ensemble model that has Naive Bayes as a base classifier and the Logistic Regression as a meta classifier model. The performance of the proposed method is compared with common machine learning models such as Naïve Bayes and Logistic Regression using the sentiment140 dataset, experiments were carried out and their accuracy was determined. The results of the experiment endorse the proposed methodology. exhibits better outcomes of attaining accuracy score of 86%.

Keywords: sentiment analysis, ensemble model, logistic regression, natural language processing, naive bayes, machine learning.

1. Introduction

Utilizing the People's feelings towards a particular entity from the user's point of view one can post his own content through various social media sites, such as his or her tweets, microblogging, or online social media sites. From one perspective, many social media sites

DOI: https://doi.org/10.36548/jtcsst.2023.2.003
Received: 02.04.2023, received in revised form: 02.05.2023, accepted: 17.05.2023, published: 27.05.2023

share application programming interfaces (APIs) that facilitate data collection and analysis by researchers and developers. However, these types of online dating have some drawbacks. It can interfere with the sentiment analysis process. The first mistake is that because people are free to post their own content, opinions cannot be guaranteed. The second error is such a basic truth. Ground truth is like a label attached to a particular person. An opinion indicating whether the opinion is positive, negative, or neutral.

Social media is flooded with millions of user-generated posts and content. Most users use social media to send and share their personal views and opinions. I feel it through pictures and words. But there lies the largest challenge.

The main goal of this paper is to classify tweets. Express a positive or negative mood. By using Natural language toolkit (NLTK) and an ensemble classifier, we can analyze the scores and classify them as positive or negative. The Lexicon in the NLTK library provides a way to analyze ratings and report results based on lexicons. Emotion expressed by a posted tweet.

A sentiment analyzer is available for them to view and analyze tweets to show how happy users are about their opinions. Validate your expertise and help correct any mistakes found to be important. And this research does work through a web interface Process comments and ensemble classifiers using NLTK, classifying them as positive and negative.

2. Literature Survey

There is an important role for sentiment analysis in the subject of text classification, and numerous researchers have looked into how sentiment analysis acts by identifying the emotions that are conve yed in the text.

This paper [2] says emotion placement and the effectiveness of the three-function coordination method were discussed in the context of Twitter sentiment analysis. Focusing on evaluating and comparing several ensembles and the three ensemble techniques, we performed a series of comparative studies using two feature set schemes, five diverse datasets, three ensemble approaches, and three ensemble strategies. The feature set type is designed for sentiment analysis. a feature set based on word relationships and a feature set based on parts of speech Then, as basis classifiers for each feature set, three well-known machine learning

algorithms, were among the three ensemble methods analysed and using the algorithms of NB, maximum entropy, SVM, bagging, boosting, k-partitioning, and biassed k-fold partition with an accuracy of 85.58%.

This paper [3] proposes a differential evolution-based learning method (DEW Vote). The proposed procedure randomly selects a base classifier. Determine the appropriate weights for each base classifier based on how confidently the DE algorithm predicts the future. In the DEW Vote approach, we acquire Naive Bayes, Bayes Net, -Nearest Neighbour (-NN), and ZeroR as our five foundation classifiers. DEW Vote shows better accuracy than the other ensemble methods except for the datasets for diabetes and the Segment Challenge, where majority voting performs better than the other ensemble techniques. With majority records, DEW Vote and Majority Voting perform better. But bagging and boosting outperform voting records.

In this paper [4], to implement the Naive Bayes algorithm, a pre-trained SentiWordNet dictionary must be available online. It consists of different collections of words with synonyms and polarities. Analysts use Hadoop frameworks to prepare a collection of movie information available on the Twitter site as reviews, input, and opinions.

This paper [5] used three state-of-the-art classifiers: naive Bayes, support vector machines, and maximum entropy. These three classifiers were trained, and the training data was then validated five times. After training, we used the trained classifier to categorise fresh tweets as positive or negative using the test dataset. The performance of several function combinations was evaluated by feeding them to various machine learning algorithms (NB, SVM, and MaxEnt). Therefore, the objective of this investigation was to evaluate how different feature combinations are utilised in sentiment analysis with an accuracy of 89%.

In this paper [6], to improve the performance of the classifier, we preprocessed the extracted data before examining it. The first breaks the input stream into individual words. Text preprocessing means removing noise such as stop words, punctuation marks, terms that are not very important in the context of the text, etc. Loading text from many different data sources and various different filtering techniques, and finally, analysing text data.

This paper [7] To enhance the efficiency and accuracy of sentiment classification techniques, ensemble classifiers are proposed that combine basic learning classifiers to form a single classifier. The results show that the proposed ensemble classifier outperforms the standalone classifier and the majority ensemble classifier using the ensemble algorithm of NB, RF, LR, and SVM with an accuracy of 75.81%.

This paper [8] approaches document-level and sentence-level sentiment analysis. The sentiment analysis approach is also represented incorrectly. Describe Twitter's sentiment analysis approach, which includes supervised, unsupervised, lexicon, and hybrid approaches. Finally, discussion and comparison of the latter were highlighted. Different types of sentiment analysis applied to the Twitter dataset were analysed. Twitter sentiment analysis using supervised ML and dictionary-based approaches to Twitter mood performance the classifier mainly depends on the number of training sessions. Data and feature sets are extracted. mood of Twitter analysis strategy based on machine learning techniques Popular, especially SVM and NB classifiers, and use the algorithms of NB and SVM with an accuracy of 82.8%.

This paper [9] extracted the number of tweets by prototyping, divided the tweet results into positive and negative, and conducted sentiment analysis. Twitter information is collected for research using the Twitter API. Two common methods that are equally used are machine learning and dictionary-based methodology. We use a dictionary-based methodology. Analyse the information and ideas posted by different people. At this point, the extreme placement of this information has been completed. For example, tweets collected after review (positive, negative, and neutral) are divided into three classes using the algorithm K-nearest neighbour, NB.

3. Materials and Methods

The proposed method for emotion classification's full architecture is presented in the current section. This study proposes a two-part strategy to categorise the feelings of tweets. It begins by applying a lexical dictionary to extract the sentiments from the tweets before classifying them as positive or negative.

A. Dataset

The dataset used for the many different types of trials in this paper, Sentiment140, was obtained from Kaggle, a public repository for benchmark datasets. 1.6 million tweets were collected from the dataset using the Twitter search API. The dataset has an even distribution of tweets, with 0.8 million positive and 0.8 million negatives. The labels for the tweets in this dataset are 0 for negative sentiment and 4 for positive sentiment, respectively.

Table 1. Sentiment of the dataset

Polarities	Sentiment
0	Negative
4	Positive

B. Data Pre-Processing

The majority of the data that is taken from internet platforms is unstructured or semi-structured, which means that it contains extraneous information that is irrelevant to the study. Pre-processing the data is therefore crucial for ridding it of redundant and noisy information. Pre-processing is therefore necessary to conserve computational ML models so they can train more successfully with fewer resources and a simpler process, which leads to more accurate predictions.

Data cleaning involves eliminating usernames, punctuation, stop words, lowercase transformation, and stemming, among other things. These stages are as follows

B. Term Frequency Inverse Document Frequency (TF - IDF)

The Inverse Document Frequency (IDF) One word is usually used. The more common its use is, the lower its score. A measure of how often a term appears in a given document is called inverse document frequency (IDF). IDF represents word meaning based on word rarity. IDF scores are higher for rare words. When calculated by the formula,

D. Machine Learning Models

This study integrates NB and LR, two ML models, to do sentiment analysis on tweets. In order to perform classification tasks, ML models go through a training and testing process

E. Naive Bayes

Naive Bayes works better with categorical input variables compared to numeric variables. This helps you forecast data by making predictions based on past results.

F. Logistic Regression

Logistic regression is a statistical technique used for building machine learning models. Calculate or predict the probability that a binary (yes or no) event will occur. One of the most commonly used methods to solve binary classification problems is logistic regression. The logistic equation, often called the sigmoid function, is how LR became famous. Each evaluated integer can be assigned a value between 0 and 1.

G. Explained Architecture of NB-LR Model

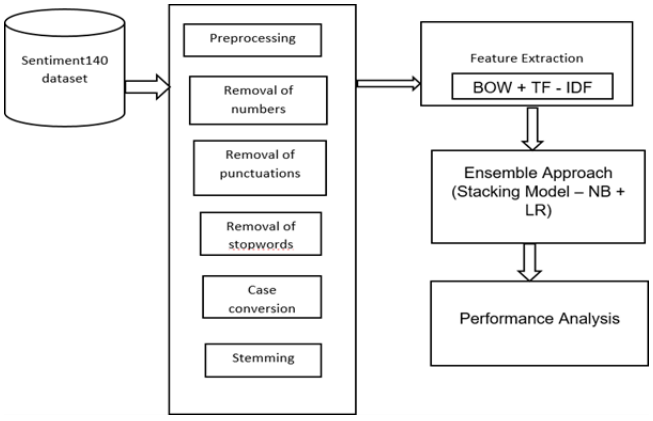

Figure 1. Architecture of proposed Model

This work incorporates a stacked approach using a combination of NB and LR to classify tweet sentiment. One of the ensembles of heterogeneous base learners and meta learners that makes final predictions using the output predictions of base learners is called stacking.

H. Proposed Method

After preprocessing, the data that has been preprocessed is split into training and testing sets in an 80:20 ratio. The recommended NB-LR model is then trained on the training set and assessed on the testing set for accuracy and confusion matrix.

4. Result and Analysis

A. Accuracy

Accuracy is a measure used to determine the best model for identifying relationships and patterns between variables. For test data, accuracy is defined as the ratio of correct predictions to all predictions. Scores for classifier accuracy vary from 0 to 1, with 1 signifying that all predictions are accurate. One of the simplest classification metrics to use is accuracy, which is calculated as the proportion of correct predictions to all other guesses.

B. Confusion Matrix

A table known as a confusion matrix is frequently used to analyze the way a classification model performs on a set of test results whose actual values can be determined. It is useful because it allows you to directly compare values such as true positives, false positives, true negatives, and false negatives. Fig.2 -4 shows the confusion matrix observed for the different machine learning Models.

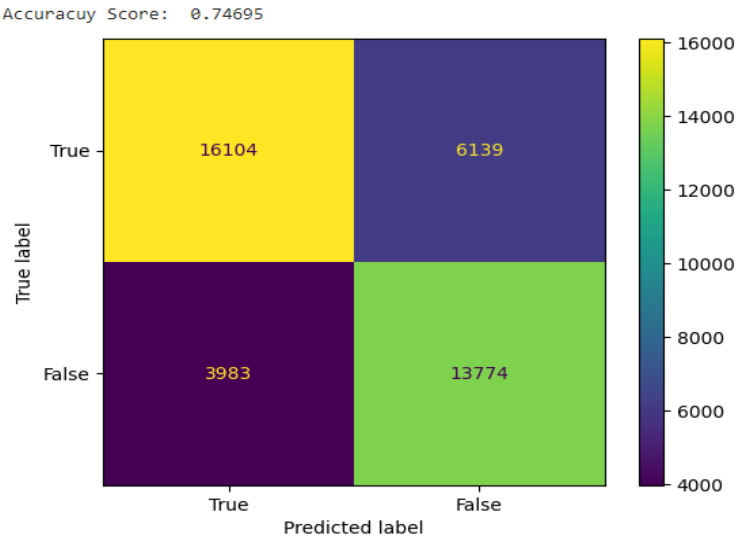

Figure 2. Confusion Matrix for Naive Bayes

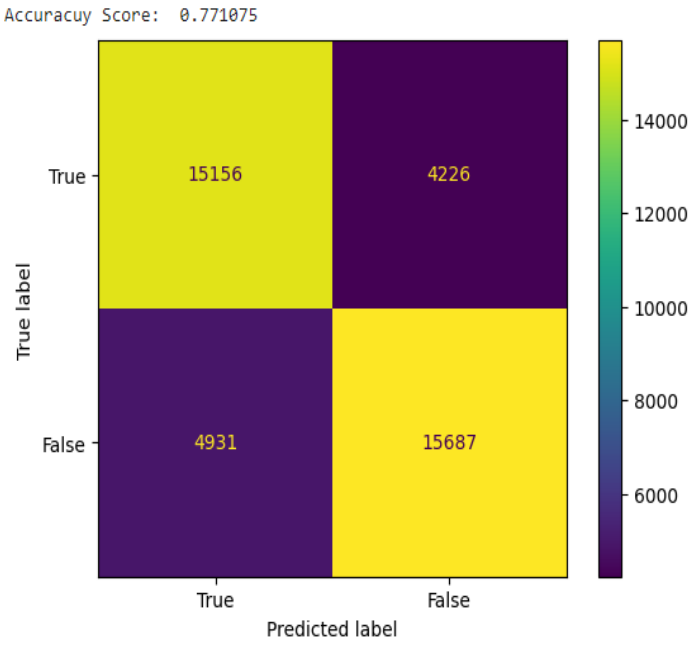

Figure 3. Confusion Matrix for Logistic Regression

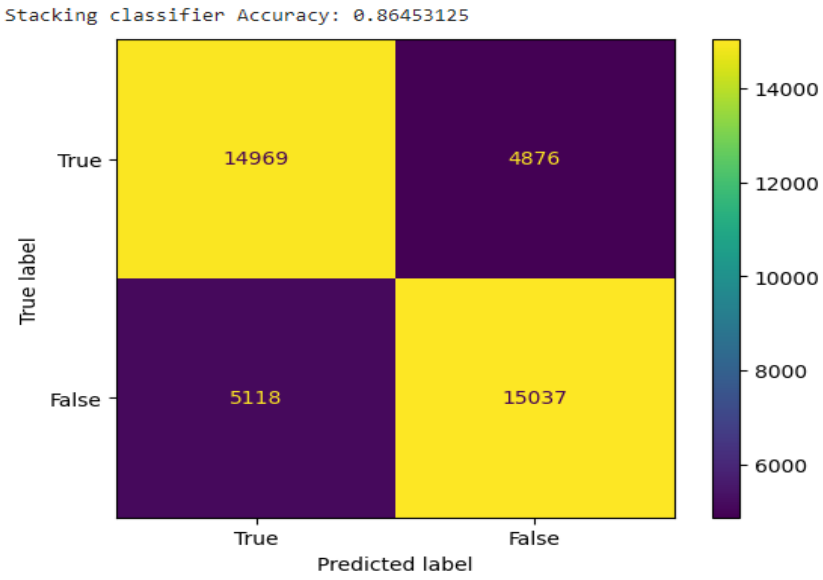

Figure 4. Confusion Matrix for Ensemble (Stacking) Classifier

Accuracy is calculated from the confusion matrix. The proposed model suggests that the Ensemble Classifier provides the largest number of correct predictions using the TF-IDF features of the Sentiment140 dataset. At this stage, tweets are classified using Logistic Regression and Naive Bayes, yielding 77% and 74% respectively. After, use an ensemble (stacking) classifier to achieve 86% accuracy.

C. Accuracy of the Classifiers

Table 2. Accuracy of the Classifiers

Classifier	Accuracy
Naïve Bayes	74%
Logistic Regression	77%
Ensemble (Stacking) Classifier	86%

5. Conclusion

The suggested method improves the sentiment analysis performance of tweets using sentiment 140 dataset by combining the ensemble and lexical dictionary approach. Features were extracted using BOW + TF-IDF. The ensemble model using the original sentiments indicated above are used to categorize tweets. The effectiveness and viability of the proposed ensemble model are demonstrated by optimal results showed accuracy of 86%which is 9% and 12% higher than Logistic Regression and Naive Bayes Classifier respectively. The ensembled model that has been proposed might be modified in the future using deep learning approach. The accuracy of the model can be increased by using pre-processing techniques like POS tagging, and the work is expanded to detect mockery, fraudulent reviews, false advertisements, spam emails, and many other things by including the word-embeddings.

References

[1] Gaye, B., Zhang, D. and Wulamu, A., 2021. A tweet sentiment classification approach using a hybrid stacked ensemble technique. Information, 12(9), p.374.

[2] R. Xia, C. Zong, and S. Li, ‘‘Ensemble of feature sets and classification algorithms for sentiment classification,’’ Inf. Sci., vol. 181, no. 6, pp. 1138–1152, Mar. 2011.

[3] Y. Zhang, H. Zhang, J. Cai, and B. Yang, ‘‘A weighted voting classifier based on differential evolution,’’ Abstract Appl. Anal., vol. 2014, pp. 1–6, May 2014.

[4] H. Parveen and S. Pandey, ‘‘Sentiment analysis on Twitter data-set using naive Bayes algorithm,’’ in Proc. Int. Conf. Appl. Theor. Comput. Commun. Technol., Jan. 2016, pp. 416–419.

[5] qbal, N.; Chowdhury, A.M.; Ahsan, T. Enhancing the performance of sentiment analysis by using different feature combinations. In Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 8–9 February 2018; pp. 1–4.

[6] V. Kalra and R. Aggarwal, ‘‘Importance of text data preprocessing & implementation in RapidMiner,’’ in Proc. 1st Int. Conf. Inf. Technol. Knowl.Manage., vol. 14, Jan. 2018, pp. 71–75

[7] Onan, A.; Korukoğlu, S.; Bulut, H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst. Appl. 2016, 62, 1–16.

[8] Alsaeedi and M. Zubair, ‘‘A study on sentiment analysis techniques of Twitter data,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, pp. 361–374, 2019.

[9] Kariya and P. Khodke, ‘‘Twitter sentiment analysis,’’ in Proc. Int. Conf. Emerg. Technol. (INCET), Jun. 2020, pp. 212–216.