Twitter Sentiment Analysing Using AI

Dr.M. Sengaliappan[1], **N.Anbarasan[2]**

[1]Head of the Department, Department of Computer Applications, Nehru College of Management, Coimbatore, TamilNadu, India ncmdrsengaliappan@nehrucolleges.com
[2] Student of II MCA, Department of Computer Applications, Nehru College of Management, Coimbatore, TamilNadu, India Anbu47123@gmail.com

ABSTRACT

Sentiment analysis is a technique used to gather viewpoints, attitudes, and feelings from social media platforms like Twitter. It is now a well-liked field of study.

Textual data is the main emphasis of the traditional approach to sentiment a Twitter is the most popular microblogging social networking site where users can send updates in the form of tweets on various topics. In this work, a labeled dataset that is freely accessible on Kaggle is employed, and a thorough set of pre-processing procedures is organized to gradually make the tweets suitable for handling using standard language handling techniques. As every instance in the collection consists of two tweets and a sentiment. Hence, machine learning under supervision is employed. Artificial intelligence classifiers can be used for this. These classifiers assess opinions on certain entities by endorsing political parties, businesses, analysts, etc. Machine learning methods for data appropriately classify the tweets through training. This means that machine learning techniques are more effective and quicker at performing sentiment analysis because this method does not require a word database.

KEYWORDS:

Supervised machine learning, Sentiment analysis ,Twitter,Data mining, Product evaluation ,ROC ,Classification ,Naive Bayes Logistic regression, Support vector machine and Linear SVC

Introduction:

In recent years, social media platforms have become invaluable sources of real-time information and public sentiment. Among these platforms, Twitter stands out due to its concise format and widespread usage, making it a rich ground for sentiment analysis. Understanding public opinion on various topics—from political events to consumer preferences—can provide significant insights for businesses, policymakers, and researchers alike.

Sentiment analysis involves the use of natural language processing (NLP) and machine learning techniques to classify text as positive, negative, or neutral. Supervised machine learning, in particular, offers a robust framework for this task by leveraging labeled datasets to train models that can accurately predict sentiments in unseen data. This approach not only improves classification accuracy but also allows for the adaptation of models to specific domains or contexts.

The objective of this study is to explore the effectiveness of supervised machine learning techniques for sentiment analysis on Twitter data. By employing various algorithms, such as Support Vector Machines, Decision Trees, and Neural Networks, we aim to identify the strengths and weaknesses of each method in capturing the nuanced sentiments expressed in tweets. Additionally, we will discuss the challenges associated with preprocessing Twitter data, such as dealing with slang, abbreviations, and the dynamic nature of language on social media.Through this research, we hope to contribute to the growing body of literature on sentiment analysis, providing insights that can enhance the predictive capabilities of machine learning models in understanding public sentiment on Twitter. Ultimately, our findings may serve as a foundation for further advancements in sentiment analysis techniques, offering practical applications for businesses, researchers, and other stakeholders interested in gauging public opinion.

RELATED WORK:

Sentiment analysis involves closely examining how one's thoughts and points of view relate to how your mentality and feelings come over in everyday language in relation to a certain situation. The primary reason for selecting Twitter's profile information is the platform's ability to provide subjective data [5]. Repeated events demonstrate sentiment analysis's amazing progress, which can control the whole spectrum of behavior and emotions for different networks and themes and outperform good versus negative. Extensive research has been conducted on the topic of sentiment analysis, employing various methodologies to anticipate social sentiments. The framework was created by Pang and Lee (2002),and it determines whether an assessment is positive or negative based on the percentage of positive terms in the total.

map-reduce functions are used in processing. They have applied naive Bayes classification using a single word [8]. The use of SA in business applications is examined in paper [9]. In addition, this article demonstrates the text analysis procedure for auditing the popular evaluation of customers toward a certain brand and offers confidential data that, once the text analysis is completed, may be used to advise decisions. Four stages of sentiment analysis have been completed in paper [10].

gathering tweets in real time up to a predetermined limit, tokenizing each one as part of pre-processing, comparing them to a word bag that is readily available, and categorizing the tweets as either positive or negative.

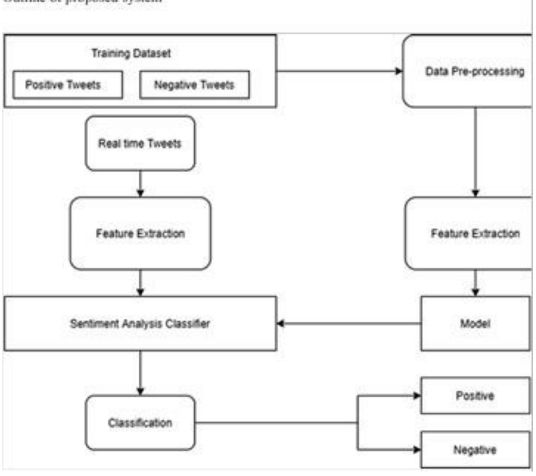
The suggested system is domain specific. There will be a user-interactive GUI where users can enter keywords.

PROPOSED SYSTEM:

The goal of the system is to analyze tweets collected from the Twitter dataset using sentiment analysis. The best algorithm has been selected after a number of algorithms have been used and evaluated against the dataset that is available. An overview of the sentiment analysis process is provided in Figure 1. The dataset will be pre-processed using the methods listed below once it has been cleaned and separated (isolated) into training and testing datasets. In order to decrease the dataset's dimension, features will be extracted. The next step is to develop a model that the classifier will use to distinguish between positive and negative tweets. The classifier will be fed real- time tweets once again in order to test the real- time data.

The procedures for managing massive incoming data are listed below:

Outline of proposed system



Datacleaning

i. Making use of different data tools to assist in cleaning the dataset.

ii. Making use of a variety of AI tools to find and remove duplicates from huge data sets.

iii. The source of mistakes should be continuously traced and monitored in order to rectify the distorted data.

iv. After the current data has been cleansed, it is validated once.

Pre-processing data:

i. Evaluating the data quality.
ii. Finding conflicting values to determine the proper data type for the features.
iii. Combine the features for improved functionality.

Clearing the Data

The format in which the data was gathered was incorrect. Information cleansing is the process of ensuring that data is accurate, dependable, and useful. Datasets typically require cleansing because they include a large amount of undesired or noisy data, also known as outliers. The presence of these anomalies could produce unsuitable outcomes.

A dataset that is far more dependable and stable is produced through data cleaning, which guarantees the elimination and improvisation of such data.

The following methods can be used to clean data:

- Keeps an eye on mistakes. Error sources and entrance points need to be continuously tracked and observed. This will assist in repairing the damaged data.
- Standardization of processes. Standardization of the point of entrance is necessary. Redundancy is minimized through standardizing the data processing procedure.

Preparing Data

Data pre-processing comes after data cleaning. It is a significant advancement in machine learning. It is the procedure that converts or encodes data into a form that a machine can comprehend. Put simply, the algorithms have no trouble understanding the properties in the dataset. A measurable attribute of the thing being seen is called a feature. A person's height, age, and gender, for instance, can be regarded as features. Every relevant tweet from Twitter will be pulled into a Twitter stream, which will have an unorganized format. Pre-handling of these unstructured tweets is necessary before applying any classifier to it. Tokenization and cleaning will be performed on the tweets beforehand.

Regarding the suggested system, the following pre-processing will be carried out:

- Making Twitter messages smaller.
- Replace a minimum of two dots with spaces.
- Replace extra spaces with a single one.
- Get rid of spaces and quotes at the end of the tweets Unigrams and bigrams are the two categories of characteristics that are taken out of the dataset.

For the features that were extracted, a frequency distribution is produced. Afterwards, the analysis is done on the top N bigrams and unigrams. Additionally, unique elements like emoticons, URLs, and user names can be found in tweets. Another element of tweets is retweets. These characteristics are not necessary for sentiment analysis to be done.

As a result, common keywords or markers like "URL," "USER_MENTION," and "EMO" are used in place of these features. Lemmatization and the elimination of stop words are once more essential procedures.

Terminate phrases. Words that are not important in search queries are known as stop words. Say, "I like to write," for instance.

ToBeEmployed Classifiers:

Bayes without learning. The result of the supervised machine learning algorithm, naive Bayes, is probability values. The naive bayes classifier is an excellent tool for handling high- dimensional issues. It makes the assumption that the odds of the various events are wholly independent of one another. A simple model called Naive Bayes [13] assigns a tweet t to class C in the following way: C =argmaxP(c|t) P(c|t)∞P(c) −n P(fi|c)

The likelihood of an occurrence By providing the occurrence of event B, an occurrence can be located. Large-scale classification problems can be solved with the Naive Bayes algorithm.

Logistic regression

Regression using logic. The result of a logistic regression is a binary prediction, such as (Y/N), (1/0), or (True/False). This also functions as a particular instance of linear regression. It results in a sigmoid, which is an S-shaped curve. Real numbers between 0 and 1 are accepted.

The logistic regression model is provided by: Output: 0 or 1.

Z = WX + B Hypothesis: hþ (x) = sigmoid(Z) 3 4 5

The goal variable in logistic regression is essentially binary. It may be able to anticipate certain target variable types. The crossvalidation estimator is employed by the logistic classifier.

Vector machine support. The approach is non-probabilistic and use a multidimensional space to represent text models as focuses.

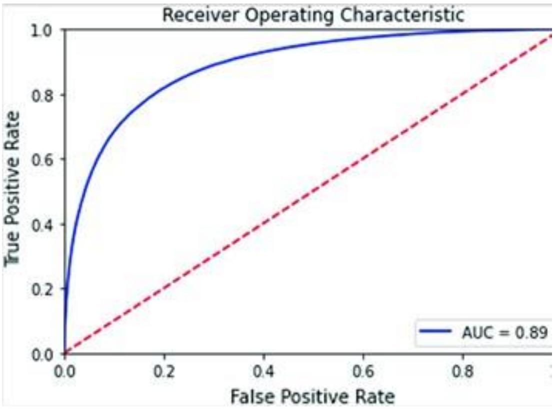
Nevertheless, logistic regression splits feature space linearly and generally performs reasonably well even when some of the variables are correlated. On the other hand, logistic regression and SVM with a linear kernel have similar performance but one may be more efficient than the other depending on the features. Logistic regression is preferred over naive Bayes for sentiment analysis because naive Bayes assumes all the features used in model building to be conditionally independent.

Ploting result

Graphs will be used to depict the results, and ROC curves will be used to compare the algorithms [14]. The true positive rate and false positive rate are plotted. The ROC curve for a MultinominalNB Model is displayed in Figure

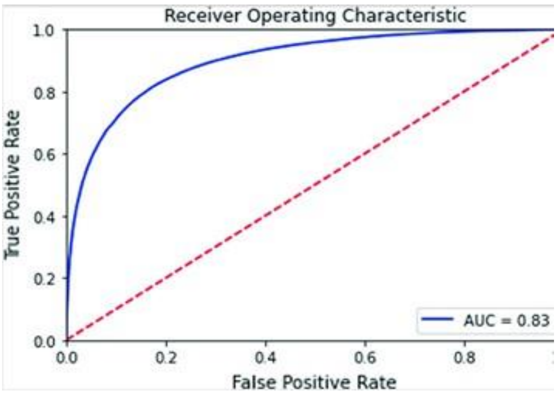
2. For the MultinominalNB Model, the Area Under the Receiver Operating Characteristics curve (AUROC) is 0.89.

Fig. 2 ROC curve for multinomial naive Bayes classifier



The ROC curve for a linear SVC model is displayed in Figure 4. For the linear SVC model, the Area Under the Receiver Operating Characteristics curve (AUROC) is 0.83.

Fig. 4 ROC curve for linear SVC



RESULT:

Sentiment140 is the dataset that was used to train the model. With 1.6 million tweets, it's a balanced dataset where 8 lakh tweets go into the positive class and the remaining 8 lakh tweets into the negative class. With a test_size of 0.20, the train_test_split technique is used to split the data. The model is trained on 12 lakh tweets, and it is tested on the 4 lakh tweets that are left behind.

Logistic Regression:

The logistic regression model's classification report is shown in Table 1. The model's accuracy is 82.47. It also demonstrates the model's recall and accuracy. Recall is the model's sensitivity, while precision is its positive predictive value [15]. Table 1

Logistic regression classification report

	Precision	Recall	F1 Score	Support
0 (Negative label)	0.81745	0.83662	0.82692	160,156
1(Positive label)	0.83236	0.81280	0.82246	159,844
Accuracy			0.82472	320,000
Macro avg	0.82490	0.82471	0.82469	320,000
Weighted avg	0.82490	0.82472	0.82470	320,000

MULTINOMINAL NAVIE BAYES:

The multinomial naive Bayes model's classification report is provided in Table 2. The model's accuracy is 80.61.

TABLE 2

Report on multinomial naive Bayes classification

	Precision	Recall	F1 Score	Support
0 (Negative label)	0.78090	0.85148	0.81466	160,156
1(Positive label)	0.83637	0.76064	0.79671	159,844
Accuracy			0.80610	320,000
Macro avg	0.80864	0.80606	0.80569	320,000
Weighted avg	0.80861	0.80910	0.80569	320,000

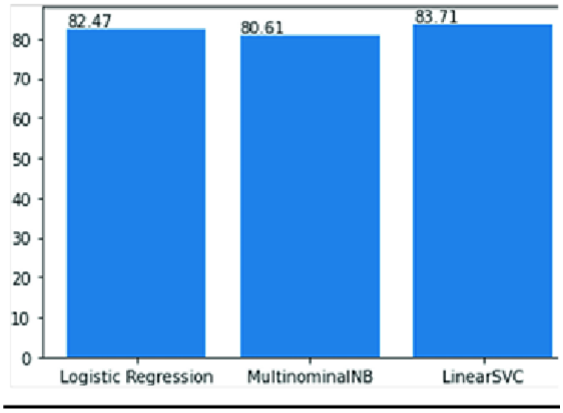
Linear Support Vector Machine :
The linear SVC model's classification report is displayed in Table 3. The model's accuracy is 83.71.

Table 3

The linear support vector machine classification report

	Precision	**Recall**	**F1 Score**	**Support**
0 (Negative label)	0.85970	0.8346	0.88889	160,01
1 (Positive label)	0.78447	0.70515	0.74270	79,989
Accuracy			0.83726	240000
Macro avg	0.82288	0.88415	0.811179	240000
Weighted avg	0.8346	0.83716	0.83716	240000

Because the dataset used for training and testing is balanced, ROC curves are useful for assessing the model's performance when the observations are evenly distributed across classes.
Since AUROC does not depend on the size of test or evaluation data, unlike accuracy, which is always influenced by test data size, it is a better indicator of classifier performance than accuracy. Moreover, because AUROC summarizes a classifier's performance in the best way possible by combining several performance metrics into a single figure. A comparison of the three sentiment analysis algorithms is presented in Figure 5; in contrast, linear SVC yields the best accuracy of 83.71, but its AUROC is lower than logistic regression's.



CONCLUSION:

This research aims to categorize a sizable corpus of twitter data into two sentiment groups: positive and negative, respectively. Sentiment features yield more accuracy than traditional text classification methods. Businesses, associations, entrepreneurs, and other entities can utilize this function to assess their goods and gain a better understanding of what customers have to say about them. Future work will involve working in additional regional languages in addition to English. In order to get the highest accuracy, it will also analyze complicated emotions like sarcasm and create a hybrid classifier.

REFERENCES

1. Neethu MS, Rajasree R (2013) Machine learning algorithms for sentiment analysis on Twitter. Third International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013; Tiruchengode, India, pp. 1–5.
2. Twitter sentiment analysis by Sebastian TM and Kumar A. Delhi Technological University's Department of Computer Engineering, Delhi, India
3. Joshi R, Tekchandani R (2016) Using supervised classifiers to analyze Twitter data comparative. In: International Conference on Computational Innovation and Technology, 2016 (ICICT). 978-1-5090-1285-5 is the ISBN.
4. Xie B, Vovsha I, Rambow O, Passonneau R, Agarwal A, Twitter sentiment analysis. Columbia University, Passonneau Department of Computer Science, New York, NY 10027, USA

5. Shamshirband S, Karim A, Moin S, Hasan A, Sentiment analysis for Twitter accounts using machine learning. Air University's Department of Computer Science, Multan Campus, Multan 60000

6. Pang B, Lee L (2008) Sentiment analysis and opinion mining

7. Go A, Bhayani R, Huang L (2009) Classifying Twitter sentiment using remote supervision. Stanford Project Report CS224N