**OPINION MINING ON TWITTER DATA USING MACHINE LEARNING**

**Mr. M. Kishore Babu1**, M.Tech (Ph. D), Assistant Professor, Department of CSE,

Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

**Vinay Rajolu2**, **Manjunath popuri3**, **Srinivas Peram4**, **Niharika Marri5**

2,3,4,5 UG Students, Department of CSE,

Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

**1 E-Mail ID:** [**kishorebabu@vvit.net**](mailto:kishorebabu@vvit.net) **(Mr. M. Kishore Babu)**

**Abstract**

The purpose of this research is to investigate how machine learning methods can be utilized for the analysis of opinions expressed on Twitter. The objective is to extract sentiment from tweets related to NASA Artemis I mission and classify them as positive, negative, or neutral. The dataset used in this research consists of tweets associated with NASA Artemis I mission collected from DATA.NASA.GOV. Preprocessing techniques such as emoji to text conversion, data cleaning, and stop words removal were applied to the dataset. The machine learning classification algorithms such as Support Vector Machine, Decision Tree, Random Forest, and Voting Classifier were applied to the preprocessed data. The performance of the algorithms was evaluated using metrics such as accuracy, precision, recall, and F1-score. The results showed that the voting classifier algorithm performed the best with an accuracy of 87.4% and precision of 0.80, indicating that machine learning techniques can effectively perform opinion mining on Twitter data. The study pinnacle the ability of machine learning techniques in extracting sentiment from NASA Artemis I tweets.

**Keywords:** Sentiment, Stopwords, TextBlob, Tweet, VADER.

**1. Introduction**

Social media platforms, including Twitter, have evolved into influential information sources for both individuals and organizations. With millions of active users sharing their thoughts and opinions on various topics, Twitter has become a valuable platform for extracting insights about public opinion. The automatic recognition and classification of the attitude or emotion conveyed in text data is referred to as opinion mining, or sentiment analysis. Opinion mining on Twitter data has gained significant attention in recent years due to a large amount of data available on the platform and its potential for various applications.

The objective of this research is to add to the increasing collection of studies related to sentiment analysis of social media data, with a specific focus on data obtained from Twitter. The results of this study can be used for drawing conclusions on NASA Artemis I mission. This paper aims to explore the potential of opinion mining on NASA Twitter data using machine learning techniques, highlighting the challenges and limitations and proposing future directions for this field.

**2. Literature Survey**

A great number of studies have looked into sentiment analysis models. The methods and models differ between fields. Only a few of the studies, though, looked into how Emoji characters were used on social media. Novak et al. [3] implemented the first Emoji lexicon. 1.6 million tweets gathered in 13 different European languages were categorized into negative, neutral, or positive categories by 83 human annotators. They discovered that about 4% of the tweets they had gathered featured emojis. They then used the sentiment score of the plain text to rank the 751 most popular Emoji characters. They used Emoji characters to represent the classes. In our work, we converted the emoji into their text format by using an emot module instead of giving scores to them. Elbagir et al. [4] removed the emoticons and corresponding emojis in the preprocessing phase, as we discussed above we can understand how the emojis are affecting the text. To improve sentiment analysis, researchers have used several different tactics. The most widely used techniques in this regard are those that rely on machine learning and lexicons. Because of their increased adaptability and precision, machine-learning techniques are becoming more and more popular among researchers. Supervised machine learning techniques are frequently used in sentiment analysis to increase the precision and effectiveness of sentiment categorization or prediction analysis. Singh et al. [5] stated that we can optimize the results of sentiment analysis by using different machine learning classifiers. They carried out their experiments by using the WEKA software tool which is very much useful for the classification of text in the text. Apporv et al. [6] stated that feature engineering with a tree kernel provides the best results instead of one-way classification. The author defines two classification models in the paper: 2-way classification models and 3-way classification models. The sentiment is classified as positive or negative in a two-way classification, and positive, negative, or neutral in a three-way classification. According to the author, tree-based kernels produced the best accuracy and feature model. Many academics have used deep learning for tweet classification and image classification [7, 8]. A Tweets Classifier for US Airline Corporations Sentiments was proposed by Rustam et al. [9]. Pre-processing was requested for the dataset by the researcher. It has been investigated how several feature extraction techniques, such as TF, TF-IDF, and word2vec, affect classification accuracy. Also, a specific dataset was used to study how long short-term memory (LSTM) was used. The researcher suggests a Voting Classifier (VC) in the study to process comparable elections. For determining results, the voting classifier must rely on spatial estimation (SE), stochastic gradient descent classifier (SGDC), and a straightforward ensemble approach. As working metrics, a variety of ML classifiers were evaluated using precision, accuracy, recall, and F1-score. According to the findings, the suggested VC is more effective than one of the phase actors. The experiment also indicated that when TF-IDF is used as a feature input, the efficiency of machine-learning students improves. According to Deepika et al. [10], good pattern recognition and model combining can boost the performance of models. The author also concludes that using feature representation approaches like Tf and Tf-IDF increases the model's accuracy.

**3. Problem Identification**

From our Literature Survey, we noticed that in the preprocessing phase emojis had been removed. Therefore, converted the emojis to text format using the emot module and the resultant text was preprocessed by NLP and preprocessed tweets are fed input to different machine learning algorithms like SVM, Decision Tree, Random Forest, and Voting Classifier.

Adding class labels to unlabeled data is one of the problems in opinion mining. Adding labels to a dataset is vital in sentiment analysis, where you get the data as reviews or comments from users, and you need to add labels to it to prepare it for sentiment analysis.

**4. Methodology**

**4.1 Data collection:**

The data was collected from DATA.NASA.GOV about NASA Artemis I mission. The collected data doesn’t have class labels. Adding labels to a dataset is very important before feeding data to supervised machine learning models to solve a problem.

**4.2 Data Preprocessing:**

The obtained tweets have to follow the below steps to process.

4.2.1 Conversion of emoji to text.

The tweets containing the emoji icons are converted to text using the emot module available from the python 3.0 version. For instance, 😀 indicates a happy and smiling state of mind so the emot module gives output as happy\_smiling\_face.

4.2.2 Conversion to lowercase.

As the machine shows the difference between uppercase and lowercase letters which have the same meaning so converting all letters to lowercase avoids such differences. For example, ‘Cat’ and ‘cat’ have the same meaning but the machine understands them as different words.

4.2.3 Removal of HTML tags and URLs.

URLs in a text point a location to the web but just like HTML tags, it doesn't provide any useful information in the tweet. These URLs and HTML tags are removed by using regular expressions. These tags won’t add any value to the text data. For example, https://www.twitter.com/vinay2301 has no meaning in tweets so these are removed.

4.2.4 Removal of Punctuations.

The reason for removing punctuations is pretty similar to lowercasing, in certain cases, we want the word wow and wow! to be treated in the exact same way. Although be careful while using punctuation, the word can't be converted to cant and can t depend upon what you set in the parameters. period, comma, apostrophe, quotation, question, exclamation, brackets, braces, parenthesis, dash, hyphen, ellipsis, colon, and semicolon are the punctuations that need to be removed from the text.

4.2.5 Removal of stopwords.

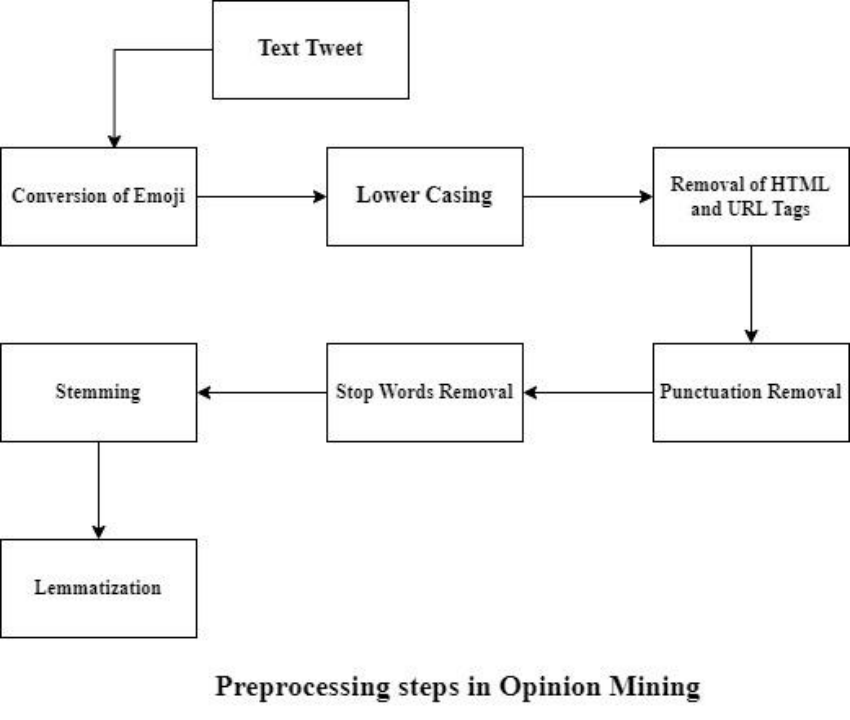
The stopwords are words such as is, are, the, so, and which present in bulk amounts in the text but they don't provide much useful information to the model, so by removing these words, we can focus on the more important information in the text.

4.2.6 Stemming

Stemming is a process of reducing the derived words to their stem or root form. Consider the tweet ‘I love running, but I hate runners’ that has running and runners which share the same root form run. After applying stemming the words are converted to run which ensures the expression in the tweet is accurate.

4.2.7 Lemmatization

Lemmatization is a linguistic process of reducing words to their base or dictionary form, which is called a lemma. In stemming some words are not converted to their actual base form to avoid this lemmatization is used so that the words are exactly converted to a root form. For example, the word ‘caring’ is converted to ‘car’ in stemming however in lemmatization ‘care’ which is the true root form.



**Fig.1** Preprocessing steps

**4.3 Class Labelling**

4.3.1 VADER:

The obtained static data doesn’t have the class labels which are very crucial in determining the sentiment of a tweet that is either positive, negative, or neutral. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool that takes the text and assigns a sentiment score [-1, 1] to each word. The overall sentiment score of the text is obtained by aggregating the individual scores of words in the text.

The compound score of the text ranges from [-1, 1].

* [-1, -0.5] indicates negative sentiment.
* (-0.5, 0.5) indicates neutral sentiment.
* [0.5, 1] indicates positive sentiment.

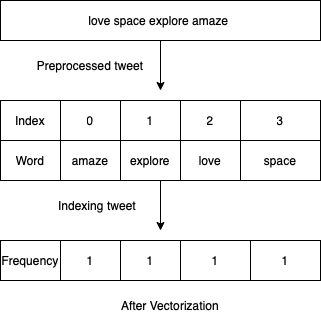
**4.4 Feature Extraction**

Feature extraction is obtaining the most relevant features or attributes from the text which are used to predict the sentiment of the given text. The quality of features obtained decides the success of the sentiment analysis system.

CountVectorizer: It converts the given text data into a matrix of token counts, which is then used as input to machine learning algorithms. It usually assigns the frequency of the word in the text.

frequency [Word] = Total number of times that word appeared in the text

For example, “I love space exploration! It’s amazing.” is a tweet. After preprocessing the tweet becomes “love space explore amaze”



**Fig.2** Feature extraction

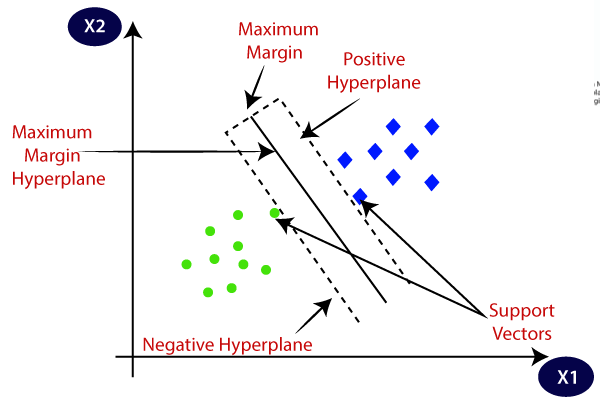
**4.5 Classification**

The extracted features are fed into machine learning algorithms which are Support Vector Machine, Decision Tree, Random Forest, and Voting Classifier and trained with a training dataset. Out of all these models, accuracy of voting classifier is highest with 87.4% and precision 0.80, SVM with 86.1%, Decision Tree with 85.6% and Random Forest with 84.1% .

**5. Implementation**

Model was developed by using supervised machine learning algorithms SVM, Decision Tree, Random Forest, and V

**5.1 Support vector Machine:** SVM is a supervised machine learning algorithm that is used to classify data points into two or more classes. It can handle high dimensional data and a large number of features effectively. It works by fitting a best hyper plane that has the highest marginal distance between the data points and plane. This plane helps to classify the tweets as positive, negative or neutral. However, SVM gave an accuracy of 86.1%.



**Fig.3** SVM hyper plane divides tweets into positive and negative.

**5.2 Decision Tree**

In the context of opinion mining, decision trees are used to classify the tweets as positive, negative or neutral. Decision tree works upon dividing the large data set into smaller subsets based on attribute selection methods and recursively performing the splitting of each sub set until the final classification is reached. The labeled dataset is used to train the decision tree algorithm and learns to classify the tweet using patterns in the training data. It has given an accuracy of 85.6%.

**5.3 Random Forest**

Random forest is an ensemble learning method that unites multiple decision trees to improve performance and reduce overfitting. It works by constructing a large number of decision trees and each decision tree is trained with a random subset of training data. Based on majority voting of all decision trees the final classification is determined. Random forest can handle high dimensional feature spaces with a large number of input features which cannot be handled by decision trees. The obtained accuracy of random forest is 84.1%.

**5.4 Voting Classifier**

A voting classifier combines the predictions of multiple machine learning algorithms to improve overall classification accuracy. In this project, the voting classifier is built by training the svm, decision tree, and random forest algorithms on the same training dataset.

Each algorithm predicts the sentiment for the given tweet and by considering the majority voting of the algorithm's final classification is made. Voting can be done in two ways one is soft voting and another is hard voting. The voting classifier produced an accuracy of 87.4% which is highest of all algorithms accuracy.

**6. Results & Conclusion**

**6.1 Results**

The original static dataset was partitioned as 70% training data and 30% testing data. The testing data was used to observe the performance of the trained model and accuracy, precision, recall, f1-score are metrics used to evaluate the model performance.

The formulas to calculate the metrics are

Accuracy = (TP+TN)/(TP+TN+FP+FN)

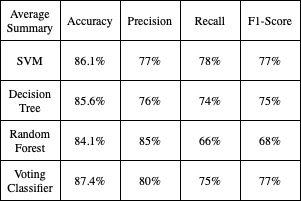
Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

F1-Score =

(2×Precision×Recall)/(Precision+Recall)

Table 1 shows the comparison of accuracy, precision, recall and f1-score of all four models.



**Table 1:** Comparison of Models

**6.2 Conclusion**

Machine learning techniques such as classification algorithms, SVM, decision trees, random forests, and voting classifiers have been widely used to analyze and classify sentiments expressed in Twitter data. The experimental results show that the voting classifier gives the highest accuracy of 87.4% which can be used to classify the new tweet into positive, negative or neutral. However, there are several limitations that must be addressed to improve the accuracy and usefulness of opinion mining on Twitter data, including limited context, and data quality. Future work in this area can focus on developing real time tools and using deep learning models can find the complex and hidden patterns between words and phrases which can improve accuracy.

**7. Limitations & Future Scope**

**7.1 Limitations**

* The considered dataset has 14110 tweets which are limited so it is difficult to classify without additional information for new context or tweet.
* The collected data contains spelling errors and abbreviations which are not preprocessed and has an effect on the accuracy of machine learning models.

**7.2 Future Scope**

* Developing real-time sentiment analysis tools can provide more immediate insights into public opinion and can help businesses and organizations respond more quickly to emerging trends or issues.
* Using deep learning algorithms such as neural networks can find the most complex relationships between words and phrases which can improve the accuracy of the sentiment analysis

**8. References**

[1] R. Katarya, A. Yadav, “ A comparative study of genetic algorithm in sentiment analysis”, In 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2018, pp. 136-141

[2] Churches, O., Nicholls, M., Thiessen, M., Kohler, M., & Keage, H. (Jan. 2014).

Emoticons in mind: An event-related potential study. Social Neuroscience, 9(2), 196–202.

[3] Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (Dec. 2015). Sentiment of emojis. PLOS One, 10(12).

[4] Elbagir, S., Yang, J. (Oct . 2019). Twitter Sentiment Analysis Based on Ordinal Regression.

[5] Jaspreet, Gurvinder, Rajinder. (2017). Optimization of sentiment analysis using machine learning classifiers. doi:10.1186/s13673-017-0116-3

[6] Apoorv Agarwal, BoyiXie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, “Sentiment Analysis of Twitter Data” Proceedings of the workshop on Language in Social Media (LSM 2011), 2011.

[7] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, ‘‘A novel stacked CNN for malarial parasite detection in thin blood smear images,’’ IEEE Access, vol. 8, pp. 93782–93792, 2020.

[8] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, ‘‘Aggression detection through deep neural model on Twitter,’’ Future Gener. Comput. Syst., vol. 114, pp. 120–129, Jan. 2021.

[9] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, ‘‘Tweets classification on the base of sentiments for US airline companies,’’ Entropy, vol. 21, no. 11, p. 1078, Nov. 2019.

[10] G. Deepika, and Dr.G. N.R. Prasad, “Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD),’’ Vol. 11, Issue 4, April 2022.