2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)

Sentiment Analysis of Tweets Using Machine Learning Algorithms

Raj Pate
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
rajpate77725@gmail.com

Siddhesh Patil
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
siddheshcp96896@gmail.com

Manthan Patil
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
patilmanthan015@gmail.com

Dr. Roshani Raut
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
roshani.raut@pccoepune.org

***Abstract*—Social media is a popular platform for individuals to express their opinions on various topics. Sentiment analysis using Twitter datasets addresses important requirements such as conducting market research, performing political analysis, and gaining social insights. It empowers businesses to derive valuable insights from user-generated content on Twitter. The objective of this research is to create a sentiment analysis model that can be applied to social media data, specifically tweets from Twitter. The sentiment analysis model is built using machine learning (ML) techniques and natural language processing (NLP) to determine if each tweet is positive or negative. This allows organizations to gauge public opinion and make informed decisions. A vast collection of tweets is used to test the model's performance, with a focus on accuracy, precision, and recall metrics. The results show that the model is successful in analyzing sentiment across various topics and domains. Out of the various machine learning algorithms examined, the Support Vector Machine (SVM) algorithm yielded the highest accuracy of 94.65%. The study as a whole emphasizes the importance of sentiment analysis in comprehending public sentiment on social media platforms.**

***Keywords*—Sentiment Analysis, Twitter, Naive Bayes, SVM, Decision Tree, NLP**

I. INTRODUCTION

The surge in social media platform usage has generated a plethora of user-created content such as text, images, and videos. The study of sentiment analysis, also called opinion mining, focuses on automatically detecting and extracting subjective information from text data. It has become a crucial area of natural language processing with a lot of research attention devoted to analyzing sentiment in social media data. Twitter, with its wide API accessibility and real-time data volume, has emerged as a popular source of data for sentiment analysis. Users of Twitter often share their opinions and emotions on various topics, making it an ideal platform for sentiment analysis. Go, R. Bhayani, L. Huang [1] offered a fundamental approach to sentiment analysis that employs algorithms, such as Naive Bayes and SVM, in conjunction with a bag-of-words model. This technique involves breaking down the text into individual words and analyzing them independently, without taking into account their sequence, to determine the underlying sentiment.

Sentiment analysis of social media data is valuable for several applications, such as monitoring public opinion, tracking brand reputation, and forecasting consumer behavior. By analyzing sentiment in social media data, researchers can gain insights into the attitudes and opinions of individuals and groups, providing useful information for decision-making in various domains. This study concentrates on sentiment analysis for social media data using a Twitter tweet database. In the paper [2], to automatically classify the feelings discovered in the tweets, classifier ensembles and lexicons are used. These techniques can help to categorize tweets as either positive or negative in response to any particular query. Sentiment analysis can be useful for organizations that want to find out public opinion about their products, as well as consumers who need products and seek opinions from others. The suggested model uses machine learning algorithms and NLP techniques to determine the sentiment present in Twitter data.

The research evaluates the model's effectiveness by measuring accuracy, precision, and recall metrics and compares it to other approaches that have been documented in the literature.

The major contribution of the work is the development of a sentiment analysis model specifically tailored for social media data, focusing on tweets from Twitter. By leveraging machine learning techniques and natural language processing, the model successfully classifies each tweet as positive or negative, allowing organizations to gain insights into public opinion. The research provides a comprehensive evaluation of the model's performance, emphasizing accuracy, precision, and recall metrics. Overall, this work contributes to the understanding and application of sentiment analysis in social media platforms, providing valuable tools for decision-making based on public sentiment.

The paper's remaining sections outline a literature review of sentiment analysis, a proposed methodology consisting of data collection, data preprocessing, feature extraction, and classification, and the results of the study, which comprise performance evaluation and comparison to existing approaches and further the conclusion of the study.

979-8-3503-2379-5/23/$31.00 ©2023 IEEE

II. Literature Survey

Sentiment analysis using Twitter data has been an active area of research for over a decade. It has been a popular research topic in the field of natural language processing and machine learning in recent years. Some of the key papers that have contributed to this field are mentioned further.

Twitter users can write short posts called tweets and share them with multiple users who use the platform [3]. As of 2016, there were 313 million active users on Twitter in a given month, of which 100 million users were active daily.

A single sentence or phrase may carry distinct meanings when used in other areas [4]. For instance, the term "unstoppable" may have a positive connotation in the context of films, plays, and similar domains, but it can have a negative sentiment when used to describe a vehicle's steering.

There are various effective methods for extracting opinions from Twitter messages [5]. However, the presence of spam and the use of widely different languages make it hard to gather opinions from Twitter.

Sarcasm poses a challenge in sentiment analysis, where a person may write something positive but intend a negative meaning, or vice versa. This complexity adds difficulty to the task of sentiment analysis [6].

Detecting sarcasm in tweets is very difficult, which is a common problem in sentiment analysis [7]. The authors presented a model that utilizes both behavioral and language-based features to identify sarcasm. The model was based on features representing different types of patterns from a text such as polarity, ambiguity, emotional content, and unexpectedness.

The data gathered from Twitter often includes spelling mistakes, symbols, URLs, and hashtags, which might lead to inaccurate results [8]. To improve the accuracy of the results, a preprocessing step is performed on the collected data. This involves removing URLs, stop words, hashtags, special symbols, emoticons, and other elements.

The data available on Twitter is not structured and is also diverse in nature containing a mix of positive and negative sentiments [9]. It also explains how Naive Bayes and SVM algorithms can be utilized to classify the types of sentiments discovered on Twitter. The various challenges associated with the methods are also clearly described in this paper. Various studies claim that logistic regression is also very effective for sentiment analysis on Twitter data [10]. The authors of the research paper put forward an unsupervised system for sentiment analysis of text documents that can ascertain the orientation of the documents based on their polarities. These systems can be used in further research.

Precision is defined as a measure that calculates the proportion of positive observations that were correctly predicted out of all the observations that were predicted as positive [11]. The recall is the proportion of positive observations that were accurately predicted out of all the observations belonging to the positive class.

Overall, these papers provide a solid foundation for conducting sentiment analysis using Twitter data, and highlight the importance of using a combination of techniques to achieve high accuracy and reliability.

In today's age of automation, machines are consistently being directed toward offering precise interpretations of individuals' expressions on social media [12]. The multifaceted perspective on the evolution of sentiment analysis by offering a comprehensive understanding of how sentiment analysis has gained prominence due to the rapid surge of abundant data available on the internet is therefore required.

In recent years, there has been a significant surge in the interest surrounding sentiment analysis in social networks. This continuous increase can be attributed to two main factors. Firstly, it can be attributed to the continuous advancements and integration of new processing techniques and domains. Secondly, it can be attributed to the inherent challenge of accurately assessing and categorizing information within various domains [13].

III. Methodology

Sentiment analysis is the process of analyzing text data to determine the overall emotional tone of a piece of content. In the context of Twitter data, sentiment analysis is often used to determine the public's opinion about a particular topic or event. Here is a methodology for conducting sentiment analysis using Twitter data.

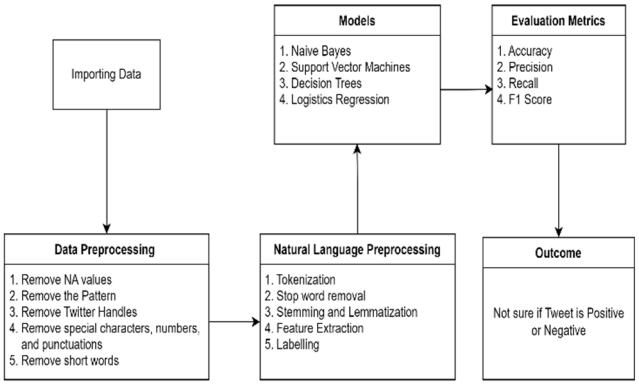

Fig 1. Research Methodology

The following Fig.1 depicts the flowchart of the research methodology.

A. Data Collection

The initial phase of Data Collection is to decide the target audience for sentiment analysis. This will help to identify the keywords and hashtags that can be used to collect data. There are various tools available for data collection from Twitter, such as Tweepy, Twitter API, or the Twitter Search API. The experimentations are done using Python programming language.

To collect data from Twitter, the search criteria need to be defined. This can be done using keywords, hashtags, or Twitter handles. The search results can also be filtered based on the date range, language, location, and other parameters. Once the search criteria are defined, the actual data collection can be started depending on requirements.

For this research, a large dataset is taken from Kaggle consisting of 31961 tweets described in Table 1.

TABLE I. DATASET DESCRIPTION

Data Set	Number of Tweets
Training Set	23624
Testing Set	5906

.

B. Data Characteristics

The study focuses on Twitter, a social networking site that is known for its unique writing format. Some of the key features of Twitter include:

- Limited tweet length: Twitter messages are short and cannot exceed 140 characters.
- High tweet volume: Twitter is considered one of the most widely used social networking sites with around 500 million tweets posted daily.
- Varied topics: On Twitter, individuals share information and opinions on a wide range of subjects, including politics, mobile technology, and more.
- Informal writing style: Tweets often contain spelling errors, slang, and emoticons, which require pre-processing for analysis.
- Emoticons: Punctuation is used to create emoticons that convey facial expressions.
- User mentions: The "@" symbol is used to mention other users in tweets.
- Hashtags: The "#" symbol is used to denote the subject or theme of a tweet.
- Retweets: The "RT" symbol is used to indicate that a tweet is a repost of another user's message.

In summary, Twitter is a commonly used platform for social networking that generates a massive volume of data on a diverse range of topics. Its unique writing style and features require specialized pre-processing techniques for analysis.

C. Data Pre-processing

Data Preprocessing is a critical step in any machine learning task, including sentiment analysis of Twitter datasets. The main goal of data preprocessing is to transform raw Twitter data into a format that is more suitable for analysis by a machine learning algorithm.

The following are some common steps in data preprocessing for sentiment analysis of Twitter datasets:

1. Data cleaning: Twitter data is often noisy and unstructured, containing irrelevant information such as emojis, hashtags, URLs, and mentions. Removing these elements can help to reduce noise and improve the quality of the data.
2. Tokenization: This involves breaking the text into smaller units called tokens, such as words or phrases. Tokenization is important because it allows the algorithm to analyze the sentiment of individual words or phrases in the tweet.
3. Stop word removal: Stop words are common words that do not carry much meaning, such as "the", "and", and "of". Removing stop words can reduce the dimensionality of the data and enhance the performance of the method.
4. Stemming and Lemmatization: These techniques involve reducing words to their base form, such as converting "running" to "run". This helps to reduce the dimensionality of the data and enhance the performance of the method.
5. Feature extraction: This involves identifying the most relevant features from the tweet, such as the presence of certain words or phrases. This step is important because it allows the algorithm to learn from the most important information in the tweet.
6. Labeling: Each tweet in the dataset needs to be labeled with a sentiment score, such as positive or negative. This can be achieved manually or using an existing sentiment analysis tool.

Overall, data preprocessing is a crucial step in sentiment analysis of Twitter datasets because it can significantly impact the accuracy and performance of the machine learning algorithm. By transforming the raw data into a more structured format, the algorithm can better identify patterns and make more accurate predictions about the sentiment of each tweet.

The word cloud displays the word frequency of the data. A word cloud is a visual depiction of a text document where words are arranged in a two-dimensional space. The size of each word in the cloud corresponds to its frequency of occurrence in the document. Larger words represent words that appear more frequently in the document [14]. The word clouds in Fig. 2 and Fig. 3 show the most frequently used preprocessed words showing positive and negative sentiments, respectively.

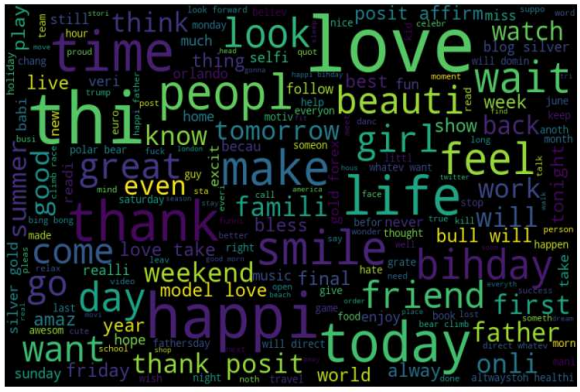

Fig 2. Positive Word Cloud

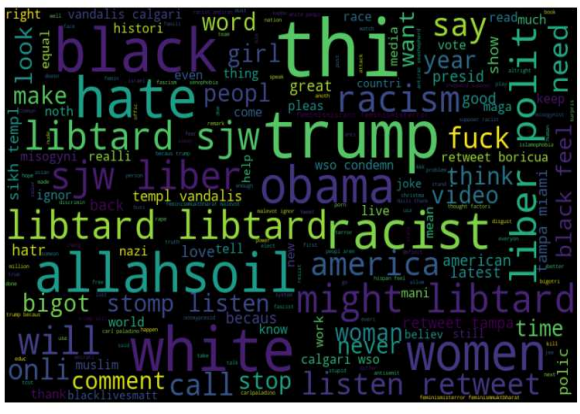

Fig 3. Negative Word Cloud

It is observed that ‘love’ is the most frequent word showing positive sentiment while ‘hate’ and 'trump' are most frequently showing negative sentiment.

D. Model Selection

Sentiment analysis is a type of natural language processing (NLP) that involves identifying and classifying subjective information in text data, such as opinions, emotions, and attitudes. There are several machine learning techniques for sentiment analysis classification.

These techniques use algorithms to learn from labeled data and then classify new data based on the patterns they find. Common machine-learning techniques for sentiment analysis include logistic regression, naive Bayes, support vector machines, and decision trees.

1. Naive Bayes

It is a commonly used machine learning approach that can be applied to analyze the sentiment of a piece of text, such as a product review or a tweet. This method utilizes Bayes' theorem to compute the likelihood of an event occurring given some evidence. When performing sentiment analysis, the algorithm can ascertain whether the text contains a positive or negative sentiment.

The Naive Bayes model is used to train the dataset which learns to associate the words and phrases in the text with the sentiment label. The Naive Bayes algorithm assumes that the features (i.e., words or phrases) in the text are conditionally independent of each other given the sentiment label. This is a simplifying assumption, but it allows the algorithm to work efficiently and effectively for many real-world sentiment analysis tasks.

Once the Naive Bayes model is trained, it can be utilized to predict the sentiment of new, unlabeled text documents. The algorithm calculates the probability of the document belonging to each of the sentiment labels (positive, negative) and selects the label with the highest probability as the predicted sentiment. Overall, Naive Bayes is a simple and efficient algorithm for sentiment analysis of Twitter datasets. It is particularly useful when the dataset is small and the features are sparse, as it can still provide reasonably accurate results even in these cases.

2. Support Vector Machines (SVM)

SVMs are generally utilized in sentiment analysis of Twitter data to analyze whether the sentiment in a tweet is positive or negative. SVMs are particularly suitable for this task because they can learn intricate patterns in high-dimensional text data and can deliver reliable predictions even with small amounts of training data.

After preprocessing the data, we can proceed to train the SVM model on the labeled dataset. The algorithm will be able to classify the data points into their respective sentiment categories based on the features provided. With the model trained, we can then use it to predict the sentiment of new, unlabeled documents. The SVM algorithm will have learned to differentiate between the different sentiment classes based on the input features.

SVMs are a widely used and successful technique for sentiment analysis of Twitter data as they possess the capability to learn intricate patterns in high-dimensional data. However, the most suitable algorithm for a particular task and data may vary based on specific data characteristics and the objective of the analysis.

3. Decision Tree

Decision trees are commonly utilized in sentiment analysis of Twitter data. The process involves training a decision tree on a labeled dataset, where each tweet is assigned a sentiment polarity label (positive or negative). The trained decision tree is then applied to classify the sentiment polarity of new, unlabeled tweets.

The decision tree algorithm operates by recursively partitioning the data based on the most significant features that differentiate the sentiment classes. The partitions are based on a criterion such as Gini impurity or information gain that evaluates the effectiveness of the split in segregating the classes. The outcome is a tree-like structure where the leaves indicate the predicted sentiment label for a given input document. One advantage of decision trees is that they can handle non-linear associations between the features and the output variable, which is valuable in sentiment analysis tasks where the link between text features and sentiment labels can be intricate. In conclusion, decision trees are a beneficial approach for sentiment analysis, particularly for tasks involving complex or non-linear associations between the text features and the sentiment labels.

4. Logistic Regression

It is a method of statistical modeling that can be used for sentiment analysis purposes. It is a binary classification algorithm that predicts one of two possible outcomes. When conducting sentiment analysis, the typical result that is expected is either a positive or negative sentiment.

The logistic regression algorithm utilizes a logistic function to establish the association between the input features and the output variable. By using the logistic function, an input value that is a real number is converted to a value that falls between 0 and 1. This resultant value represents the probability of a positive sentiment label. Logistic regression is a straightforward and comprehensible algorithm that can manage non-linear connections between the input features and the output variable. Nevertheless, in complex datasets or cases with numerous features to consider, logistic regression may not perform as well as more intricate algorithms such as SVMs.

Every machine learning technique has its own set of benefits and drawbacks, and the selection of an appropriate algorithm will be dependent on the specific requirements of the project. It is generally advised to experiment with various algorithms and evaluate their performance to determine the most effective one.

E. Evaluation Metrics

One of the commonly used and suitable methods for evaluating a classifier is the confusion matrix. The confusion matrix is a table that represents the performance of a classifier by comparing predicted and actual values. A more general form of the confusion matrix is shown in Table 2.

TABLE II. GENERAL CONFUSION MATRIX

	Predicted class 1	Predicted class 2
Actual class 1	True Positive	False Negative
Actual class 2	False Positive	True Negative

This method enables us to obtain overall assessment measures, and they encompass the following parameters [15]:

- *Accuracy*: It is a metric used to measure the overall accuracy of a classifier. It shows how accurately the classifier has predicted the result. The formula for the accuracy of a classifier is given in Eq. 1.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Predictions} \quad (1)$$

- *Precision*: It is a metric used to measure the accuracy of a model's positive predictions. In simpler terms, precision refers to the accuracy of the positive predictions made by the model, which means the proportion of correct positive predictions out of all the positive predictions made. The formula for precision is given in Eq. 2.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

- *Recall*: It is a metric used to assess a model's capacity of recognizing all positive examples. More specifically, it represents the proportion of accurate positive predictions made by the model out of all real positive instances present in the data. The formula for the recall is given in Eq. 3.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

- *F1-score*: It is a metric used to measure a model's accuracy that combines both precision and recall into a single score. This score is derived by computing a weighted average of precision and recall, with a maximum score of 1 and a minimum score of 0. The F1 score is used to assess the model's performance and determine the optimal balance between precision and recall. It is given in Eq. 4.

$$F1\ Score = \frac{2*Precision*Recall}{Precision + Recall} \quad (4)$$

IV. Results and Discussion

A. Visualisation

The bar graphs in Fig 4. and Fig 5. depict the top 10 Hashtags that are used to show positive and negative sentiments in tweets of Twitter data respectively.

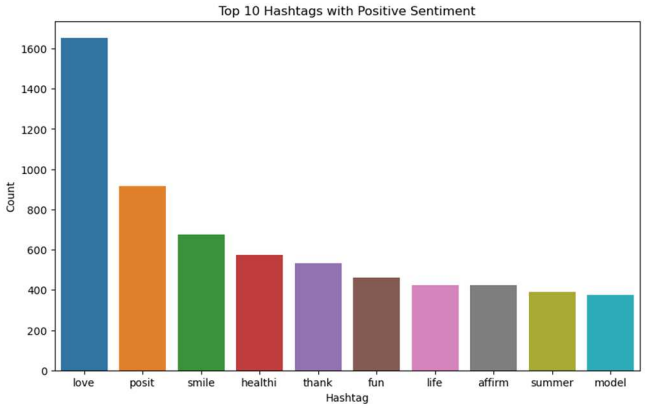

Fig 4. Top 10 Hashtags Having Positive Sentiment

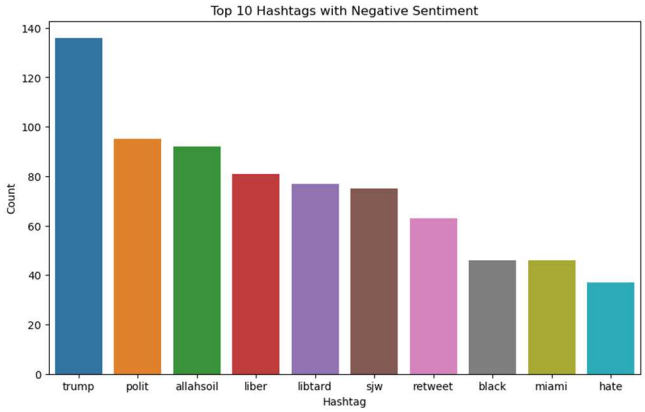

Fig 5. Top 10 Hashtags Having Negative Sentiment

The classification experiments were conducted on the Twitter dataset. The Twitter dataset contains 31961 rows. Out of them, 29530 samples were used for analysis with 23624 training samples and 5906 testing samples.

B. Accuracy Measures

The study aimed to analyze the sentiment of tweets using various machine learning algorithms on a Tweets dataset extracted using the Twitter API. The dataset consisted of a large collection of tweets, labeled with their corresponding sentiment (positive or negative). Table 3. displays the performance of different Machine Learning models, such as SVM, Naive Bayes, Decision Trees, and Logistic Regression, in sentiment analysis. The performance is evaluated based on evaluation metrics like accuracy, precision, recall, and F1-Score.

TABLE III. Comparison of Performance of Various Models

Models	Evaluation Metrics			
	Accuracy	*Precision*	*Recall*	*F1-Score*
SVM	0.9465	0.7171	0.3899	0.5052
Naive Bayes	0.9371	0.5541	0.5223	0.5377
Decision Tree	0.9246	0.4652	0.5152	0.4889
Logistic Regression	0.9433	0.6157	0.5044	0.5545

Table 4. shows the accuracy of the study conducted through various machine learning algorithms.

TABLE IV. ACCURACY EVALUATION USING VARIOUS ALGORITHMS

Training Data	Testing Data	Using SVM	Using Naive Bayes	Using Decision Tree	Using Logistic Regression
23624	5906	94.65	93.71	92.46	94.33

Fig. 6. illustrates the evaluation metrics of various models on the validation set.

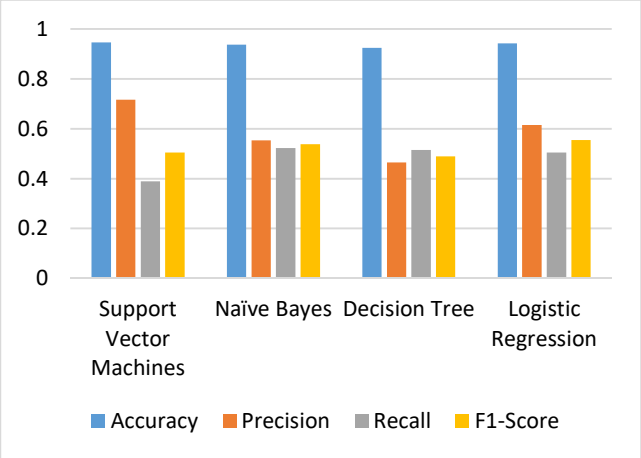

Fig 6. Evaluation Metrics of Classifiers on Validation Data

The results revealed that all the tested machine learning algorithms achieved reasonably good performance in the sentiment analysis of tweets. The accuracy scores of the models were consistently high, indicating their ability to correctly classify the sentiment of tweets.

When comparing the algorithms, it was observed that Support Vector Machines (SVM) demonstrated the highest accuracy among the tested models with a score of 94.65%. This suggests that SVM is a robust algorithm for sentiment analysis of tweets.

Furthermore, precision, recall, and F1-Score were examined to evaluate the models' performance in classifying positive, and negative tweets. The SVM algorithm provides the highest precision of 71.71% which may be very important depending on the application. The Naïve Bayes algorithm provided the highest recall of 52.23%. The F1-Score of logistic regression was relatively highest with a score of 55.45%, although the F1-Score of all algorithms used was relatively similar.

V. CONCLUSION

In conclusion, this research paper presents a system that has been developed for analyzing Twitter datasets for sentiment analysis in particular. The accuracy provided by different models suggests that they were successful in predicting the outcomes which were analyzing the sentiments of the users using the given dataset. The implications of these findings are significant, especially the growing number of young people prevalent on various social media platforms. The research conducted serves as a basis for further studies in this field of classification problems and highlights the need for continued research and development to improve the working of the models.

It is generally used in companies to analyze the sentiment of tweets related to their products or services to gain insights into customer satisfaction, identify areas for improvement, and make data-driven decisions to enhance their offerings. It is found that the Sentiment analysis can be subjective as different individuals may interpret the sentiment of a tweet differently. Moreover, bias in training data or model algorithms can lead to biased sentiment analysis results, impacting the reliability and fairness of the analysis.

The future scope is to improve the accuracy of the models through an analysis based on deep learning techniques such as neural networks. Neural networks are capable of recognizing and interpreting patterns from large amounts of data, making them suitable for sentiment analysis. The use of advanced neural networks such as CNNs and RNNs has shown the potential in improving the accuracy of sentiment analysis models. Further, the future goal is to provide real-time sentiment analysis. Real-time sentiment analysis is crucial as the number of tweets and their speed continues to rise. This includes improving algorithms and techniques for analyzing large amounts of data quickly, allowing sentiment analysis models to provide instant insights.

REFERENCES

[1] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N project report, Stanford* 1.12 (2009): 2009.

[2] Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. "Tweet sentiment analysis with classifier ensembles." *Decision support systems* 66 (2014): 170-179.

[3] Alsaeedi, Abdullah, and Mohammad Zubair Khan. "A study on sentiment analysis techniques of Twitter data." *International Journal of Advanced Computer Science and Applications* 10.2 (2019).

[4] Peddinti, Viswa Mani Kiran, and Prakriti Chintalapoodi. "Domain adaptation in sentiment analysis of Twitter." *Proceedings of the 5th AAAI Conference on Analyzing Microtext*. 2011.

[5] ZhunchenLuo, Miles Osborne. "TingWang, An effective approach to tweets opinion retrieval." *Springer Journal onWorldWideWeb* (2013).

[6] Chakraborty, Koyel, Siddhartha Bhattacharyya, and Rajib Bag. "A survey of sentiment analysis from social media data." *IEEE Transactions on Computational Social Systems* 7.2 (2020): 450-464.

[7] Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. "From humor recognition to irony detection: The figurative language of social media." *Data & Knowledge Engineering* 74 (2012): 1-12.

[8] Gupta, Bhumika, et al. "Study of Twitter sentiment analysis using machine learning algorithms on Python." *International Journal of Computer Applications* 165.9 (2017): 29-34.

[9] Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of Twitter data: a survey of techniques." *arXiv preprint arXiv:1601.06971* (2016).

[10] Sharma, Richa, Shweta Nigam, and Rekha Jain. "Opinion mining of movie reviews at the document level." *arXiv preprint arXiv:1408.3829* (2014).

[11] Jain, Anurag P., and Vijay D. Katkar. "Sentiments analysis of Twitter data using data mining." *2015 International Conference on Information Processing (ICIP)*. IEEE, 2015.

[12] Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni-Hssane. "A comprehensive survey on sentiment analysis: Approaches, challenges and trends." *Knowledge-Based Systems* 226 (2021): 107134.

[13] Rodríguez-Ibánez, Margarita, et al. "A review on sentiment analysis from social media platforms." *Expert Systems with Applications* (2023): 119862.

[14] Diyasa, I. Gede Susrama Mas, et al. "Twitter sentiment analysis as an evaluation and service base on python textblob." *IOP Conference Series: Materials Science and Engineering*. Vol. 1125. No. 1. IOP Publishing, 2021.

[15] Nguyen, Heidi, et al. "Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches." *SMU Data Science Review* 1.4 (2018): 7.