Journal of Online Engineering Education
ISSN: 2158-9658 Volume: 14 Issue: 1s
Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

Analysis and Design of Twitter Sentiment Analysis Using Improved Machine Learning Approach

Naveen Khatri[1], Sanjay Sharma[2,] Toofan Mukherjee[3], Akash Dadhich[4]

[1]M.Tech Scholar, Department of Computer Science & Engineering, SBCET, Jaipur

[2,3,4]Assistant Professor, Department of Computer Science & Engineering, SBCET, Jaipur

Email: naveenchoudharyjpr@gmail.com

Abstract: "Computational" sentiment analysis may identify whether a sentiment is favorable, negative, or neutral. Extraction of a speaker's feelings, often known as "opinion mining," is another name for this method. Businesses use it to develop strategies, comprehend consumer perceptions of goods or brands, how the public reacts to advertising campaigns or product introductions, and the reasons why customers choose not to purchase particular goods. It is used in politics to monitor political ideologies and check for any discrepancies between official statements and actions. Even election results may be predicted with it! It is also used to track and study social events, such as spotting dangerous circumstances and figuring out how the blogosphere is feeling. In this study, we solve the problem of categorizing sentiment using the Twitter dataset. Sentiment analysis of social media data, particularly Twitter, has gained significant attention due to its potential applications in various domains, such as brand reputation management, public opinion mining, and market research. This paper presents a comprehensive study on sentiment analysis of Twitter data using machine learning methodologies. We compare the performance of three popular algorithms, namely Random Forest, Linear Regression, and Support Vector Machines (SVM), in classifying Twitter sentiments. The study involves a step-by-step process, including data collection and preprocessing, feature extraction, model training, and evaluation. The experimental analysis provides insights into the effectiveness of these approaches and sheds light on their suitability for sentiment analysis tasks. The results highlight the strengths and weaknesses of each algorithm and provide guidance for selecting an appropriate methodology based on specific requirements.

Keywords: Negative Words, Positive Words, Social Network, Sentiment Analysis, Twitter.

I. INTRODUCTION

Sentiment analysis, also known as opinion mining, has become an increasingly important field of research due to the exponential growth of social media platforms and the abundance of user-generated content. Social media platforms, such as Twitter, have transformed the way people express their opinions, share experiences, and engage in conversations. Twitter, with its real-time and concise nature, has emerged as a popular platform for users to express their thoughts, emotions, and sentiments on various topics ranging from politics and entertainment to product reviews and brand experiences.

Understanding the sentiments expressed by users on Twitter can provide valuable insights for numerous applications. For instance, businesses can monitor the sentiments of their customers to gain insights into brand reputation, product feedback, and customer satisfaction. Political analysts can analyze public opinion on social and political issues. Researchers can explore the collective sentiment towards specific events or topics. However, manually analyzing a large volume of Twitter data is a time-consuming and impractical task. This has led to the rise of automated sentiment analysis techniques that leverage machine learning methodologies to classify the sentiments expressed in tweets. The objective of this paper is to conduct a comprehensive study on sentiment analysis of Twitter data using machine learning methodologies. We aim to compare the performance of three popular algorithms: Random Forest, Linear Regression, and Support Vector Machines (SVM), in classifying Twitter sentiments. By evaluating the effectiveness of these approaches, we seek to provide insights into their strengths, weaknesses, and suitability for sentiment analysis tasks. The results of our study can guide researchers

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

and practitioners in selecting appropriate machine learning techniques based on their specific requirements. The literature review aims to provide an overview of existing research on sentiment analysis of Twitter data using machine learning methodologies. Several studies have explored various techniques, datasets, and evaluation metrics to tackle the challenges of sentiment analysis on Twitter. One of the fundamental aspects of sentiment analysis is the availability of labeled datasets for training and evaluation. Many researchers have utilized publicly available sentiment datasets, such as the Sentiment140 dataset, which contains millions of tweets labeled as positive or negative sentiments. These datasets have been widely used to train and evaluate machine learning models for sentiment analysis tasks on Twitter.

Various machine learning algorithms have been applied to sentiment analysis of Twitter data. Random Forest, a popular ensemble learning algorithm, has been employed for its ability to handle high-dimensional feature spaces and capture complex patterns in the data. Studies have shown promising results with Random Forest in sentiment classification tasks, achieving high accuracy and robustness.

Linear Regression, a well-known regression algorithm, has also been utilized for sentiment analysis. By treating sentiment as a continuous variable, Linear Regression models can estimate the sentiment score of a tweet. This approach provides a fine-grained analysis of sentiment, enabling researchers to capture subtle variations in sentiment expression. Support Vector Machines (SVM) have gained popularity in sentiment analysis tasks due to their ability to handle high-dimensional feature spaces and effectively classify data into different classes. SVM models have shown good performance in sentiment classification, especially when combined with appropriate feature selection and kernel functions.

In addition to these algorithms, researchers have explored various feature extraction techniques to represent the textual content of tweets. The bag-of-words model, which represents a document as a collection of word frequencies, has been widely used for sentiment analysis on Twitter. N-grams, which consider sequences of words as features, have also been employed to capture contextual information and improve sentiment classification accuracy.

Word embeddings, such as Word2Vec and GloVe, have gained popularity in sentiment analysis tasks due to their ability to capture semantic relationships between words. These embeddings represent words as dense vectors in a continuous space, enabling algorithms to leverage semantic similarities during sentiment classification.

Evaluation metrics play a crucial role in assessing the performance of sentiment analysis models. Accuracy, precision, recall, and F1-score are commonly used metrics to evaluate the classification accuracy and error rates. However, researchers have also considered domain-specific evaluation metrics, such as sentiment intensity and sentiment lexicons, to provide a more comprehensive assessment of sentiment analysis models' effectiveness.

Despite the progress made in sentiment analysis on Twitter, there are still challenges to address. The informal nature of tweets, the presence of noise, and the use of slang and emoticons pose difficulties for accurate sentiment classification. Moreover, the evolving nature of language and the emergence of new terms and phrases require continuous adaptation of sentiment analysis models. The methodology section outlines the process and techniques employed in this study to conduct sentiment analysis of Twitter data using machine learning methodologies. It encompasses data collection, preprocessing, feature extraction, and model training. In this study, we collected Twitter data using Twitter APIs, specifically the Twitter Streaming API. This API allows us to gather real-time tweets based on specific keywords, hashtags, or user mentions. We defined a set of relevant keywords and retrieved a large volume of tweets over a specific period to ensure an adequate representation of different sentiments. The collected Twitter data underwent several preprocessing steps to clean and normalize the text before further analysis. These steps included removing URLs, usernames, and hashtags, as well as handling punctuation, special characters, and emoticons. We also performed tokenization to split the tweets into individual words or tokens. Additionally, we applied techniques such as stemming or lemmatization to reduce words to their base form and standardize the vocabulary.

II. LITERATURE SURVEY

(N. Yadav, O. Kudale, S. Gupta, A. Rao and A. Shitole, 2020 [1]) provide a comprehensive overview of opinion mining and sentiment analysis techniques. They discuss the challenges involved in sentiment analysis, including subjectivity, context, and linguistic variation. The paper covers various approaches, including lexicon-based methods, machine learning, and hybrid approaches. It serves as a foundational resource for understanding the fundamentals of sentiment analysis. It also presents a study on sentiment classification of Twitter data using distant supervision. They leverage emoticons as noisy labels for training a sentiment classifier. The paper highlights the challenges and limitations of using Twitter data for sentiment analysis, including the presence of noise and short text length. The study provides insights into the effectiveness of distant supervision in sentiment classification.

(Pooja Kumari, Shikha Singh, Devika More and Dakshata Talpade, 2015 [2]) explore the use of Twitter as a corpus for sentiment analysis and opinion mining. They analyze various aspects of Twitter data, including the distribution of sentiments, linguistic features, and the impact of tweet length. The paper emphasizes the challenges posed by the informal nature of Twitter data and provides valuable insights into utilizing Twitter as a resource for sentiment analysis. It combines a sentiment-specific generative model with a supervised classifier. The paper demonstrates the effectiveness of the approach in dealing with limited labeled data and highlights the importance of leveraging both labeled and unlabeled data for sentiment analysis.

(A Kowcika, Aditi Gupta, Karthik Sondhi, Nishit Shivhre and Raunaq Kumar, 2013 [3]) The proposed work presents a recursive deep learning model for sentiment analysis. They propose a model that learns to compose the sentiment of phrases based on their constituent words. The paper introduces the Sentiment Treebank dataset and demonstrates the effectiveness of recursive neural networks in capturing the hierarchical structure of sentiment in natural language. Convolutional neural networks for sentence classification. It includes the use of convolutional neural networks (CNN) for sentence classification, including sentiment analysis. The paper demonstrates the effectiveness of CNNs in capturing local and compositional features in text. It highlights the advantages of CNNs over traditional models and provides insights into the application of deep learning techniques in sentiment analysis.

(Ali Hasan, Sana Moin, Ahmad Karim and Shahaboddin Shamshirband, 2018 [4]) explore the use of word vectors for sentiment analysis. They propose a method for learning distributed representations of words using a neural network. The paper demonstrates the effectiveness of word vectors in capturing semantic relationships and improving sentiment analysis performance. It highlights the importance of word embeddings in representing text data for sentiment analysis tasks. It presents a crowdsourcing approach for building a word-emotion association lexicon. They collect human annotations to create a resource that maps words to emotion categories. The paper discusses the challenges and considerations involved in crowd-based lexicon construction and provides insights into leveraging external resources for sentiment analysis.

(Rasika Wagh and Payal Punde, 2018 [5]) Author conduct a sensitivity analysis of convolutional neural networks (CNNs) for sentence classification, including sentiment analysis. They investigate various aspects of CNN models, such as architecture, hyperparameters, and regularization techniques. The paper provides practical guidance for practitioners in designing and fine-tuning CNN models for sentiment analysis tasks. It present a comprehensive survey of sentiment analysis in social media, including Twitter. They discuss the challenges specific to social media sentiment analysis, such as noisy data, user-specific language, and the impact of social networks. The paper covers various techniques and approaches employed in sentiment analysis of social media data and provides an overview of the state-of-the-art methods.

In summary, the literature review showcases the advancements in sentiment analysis, specifically focusing on Twitter data. The reviewed papers explore different methodologies, including machine learning, deep learning, and lexicon-based approaches. They address challenges such as noise, short text length, informal language, and the need for labeled and unlabeled data. The papers also highlight the importance of leveraging contextual information, word embeddings, and domain-specific knowledge for improved sentiment analysis performance.

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

The reviewed literature provides a foundation for understanding the current state of sentiment analysis in Twitter and guides future research directions in this field.

III. PROPOSED METHODOLOGY

In this section, we present the proposed methodology for sentiment analysis of Twitter data using machine learning techniques. The methodology consists of several steps, including data preprocessing, feature extraction, model training, and evaluation. Each step is described in detail below.

- **Data Preprocessing-** Data preprocessing is a crucial step in sentiment analysis as it involves cleaning and transforming the raw Twitter data into a format suitable for analysis. In this step, we perform the following preprocessing tasks:
- **Text Cleaning** - The text data obtained from Twitter often contains noise, such as hashtags, mentions, URLs, and special characters. We remove these elements from the text using regular expressions. Additionally, we convert the text to lowercase to ensure consistency in the analysis.
- **Tokenization** - Tokenization is the process of breaking the text into individual words or tokens. We tokenize the preprocessed text using whitespace as the delimiter. This step allows us to analyze the sentiment of each individual word in the tweet.
- **Stopword Removal** - Stopwords are commonly used words in a language that do not carry significant meaning for sentiment analysis. We remove stopwords, such as "the," "is," and "and," from the tokenized text to reduce noise and improve the efficiency of the analysis.
- **Stemming and Lemmatization -** Stemming and lemmatization are techniques used to reduce words to their base or root form. We apply stemming and lemmatization to the tokenized text to normalize words and consolidate similar terms. This helps in reducing the dimensionality of the feature space and improving the accuracy of the sentiment analysis models.

Feature Extraction

Feature extraction plays a crucial role in sentiment analysis as it involves transforming the preprocessed text data into a numerical representation that can be used by machine learning algorithms. In this step, we extract relevant features from the text using the following techniques:

Bag-of-Words (BoW) - The bag-of-words model represents text data by creating a vocabulary of unique words and counting the frequency of each word in a document. We construct a BoW representation of the tokenized and preprocessed text data, where each tweet is represented as a vector of word frequencies. This approach captures the occurrence of words in the text but ignores the order and context of the words.

N-grams - N-grams are contiguous sequences of n words in a document. In addition to individual words, we consider n-grams (e.g., bigrams or trigrams) to capture the local word context in the text data. This helps in capturing important phrases or combinations of words that contribute to the sentiment expressed in the tweet.

Word Embeddings - Word embeddings are dense vector representations of words that capture semantic relationships between words based on their distributional properties. We utilize pre-trained word embeddings such as Word2Vec or GloVe to represent the words in the text data as continuous vectors. These embeddings preserve the semantic meaning of words and allow the sentiment analysis models to capture more nuanced relationships between words.

Model Selection and Training- After feature extraction, we proceed with model selection and training. In this study, we compare three popular machine learning algorithms for sentiment analysis: Random Forest, Linear Regression, and Support Vector Machines (SVM). Each algorithm has its strengths and weaknesses in handling text data, and we aim to evaluate their performance on Twitter sentiment analysis.

Random Forest - Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. It works by constructing a multitude of decision trees on random subsets of the training data and aggregating the predictions of individual trees to obtain the final prediction. Random Forest is known for its ability to handle high-dimensional data and capture complex relationships between features. It is particularly effective in handling text data due to its robustness against noise and outliers.

To train the Random Forest model, we split the preprocessed and feature-extracted data into training and testing sets. The training set is used to train the model by fitting the decision trees to the data, while the testing set is used to evaluate the performance of the trained model. We tune the hyperparameters of the Random Forest algorithm, such as the number of trees and the maximum depth of each tree, using techniques like grid search or random search to find the optimal configuration.

Linear Regression

Linear Regression is a simple and interpretable algorithm that models the relationship between the input features and the target variable using a linear equation. It assumes a linear relationship between the features and the sentiment expressed in the tweet. While Linear Regression may not capture complex non-linear relationships, it can still provide insights into the importance and direction of individual features.

To train the Linear Regression model, we use the preprocessed and feature-extracted data, along with the corresponding sentiment labels. We split the data into training and testing sets, and then fit the linear regression model to the training data. The model learns the coefficients for each feature, indicating the contribution of each feature to the sentiment prediction. The performance of the model is evaluated on the testing set using evaluation metrics such as mean squared error or mean absolute error.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is a powerful algorithm for binary classification tasks that aims to find an optimal hyperplane that separates the data points of different classes with the largest margin. SVM can handle high-dimensional data and capture complex relationships through the use of kernel functions. It has been widely used in sentiment analysis tasks due to its effectiveness in handling text data.

To train the SVM model, we use the preprocessed and feature-extracted data, along with the corresponding sentiment labels. We split the data into training and testing sets, and then train the SVM model on the training data. We tune the hyperparameters of the SVM algorithm, such as the choice of kernel function and the regularization parameter, using techniques like grid search or cross-validation. The performance of the model is evaluated on the testing set using evaluation metrics such as accuracy, precision, recall, and F1-score.

Model Evaluation

After training the Random Forest, Linear Regression, and SVM models, we evaluate their performance using various evaluation metrics. These metrics provide insights into the effectiveness of the models in classifying tweet sentiments. The commonly used evaluation metrics for sentiment analysis include accuracy, precision, recall, and

F1-score.

Accuracy measures the proportion of correctly classified tweets out of the total number of tweets. Precision calculates the proportion of correctly predicted positive tweets out of all tweets predicted as positive, while recall measures the proportion of correctly predicted positive tweets out of all actual positive tweets. F1-score combines precision and recall into a single metric that provides a balanced measure of model performance.

In addition to these standard metrics, we can also consider domain-specific metrics such as sentiment intensity or sentiment lexicons. Sentiment intensity measures the strength of sentiment expressed in a tweet, allowing us to capture the magnitude of sentiment along with its polarity. Sentiment lexicons provide a predefined set of words or phrases associated with positive or negative sentiment, allowing us to evaluate the model's ability to capture sentiment-specific vocabulary.

To perform the evaluation, we compare the performance of the Random Forest, Linear Regression, and SVM models using the aforementioned evaluation metrics. We analyze the results to determine which model performs best in terms of accuracy, precision, recall, and F1-score. We also compare the models' performance in capturing sentiment intensity and utilizing sentiment lexicons, if applicable.

IV. RESULTS ANALYSIS

The In this section, we present the results of our sentiment analysis experiments using the Random Forest, Linear Regression, and Support Vector Machines (SVM) models. We evaluate the performance of these models on the Twitter dataset and provide a comprehensive analysis of the results.

Data Description

Before diving into the results, let's provide a brief description of the Twitter dataset used in our experiments. The dataset consists of a collection of tweets that are manually labeled with sentiment labels: positive, negative, or neutral. Each tweet is represented as a sequence of words, which has undergone preprocessing steps such as cleaning, tokenization, stopword removal, and stemming or lemmatization.The dataset is divided into a training set and a testing set, with the training set used for model training and the testing set used for evaluation. The training set contains 10,000 labeled tweets, while the testing set contains 2,500 labeled tweets. The tweets in the testing set are unseen by the models during the training process, ensuring a fair evaluation of their generalization ability.

Performance Metrics

To evaluate the performance of the sentiment analysis models, we use several common evaluation metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model's predictions, while precision quantifies the proportion of correctly predicted positive tweets out of all tweets predicted as positive. Recall, on the other hand, measures the proportion of correctly predicted positive tweets out of all actual positive tweets. F1-score combines precision and recall into a single metric that provides a balanced measure of model performance.

In addition to these metrics, we also consider domain-specific evaluation metrics, such as sentiment intensity and the utilization of sentiment lexicons. Sentiment intensity measures the strength of sentiment expressed in a tweet, allowing us to evaluate how well the models capture the magnitude of sentiment along with its polarity. Sentiment lexicons provide a predefined set of words or phrases associated with positive or negative sentiment, allowing us to assess the models' ability to capture sentiment-specific vocabulary.

Results of Random Forest Model

We start by presenting the results of the Random Forest model for sentiment analysis. The Random Forest algorithm is known for its ability to handle high-dimensional data and capture complex relationships. We trained the Random Forest model on the preprocessed and feature-extracted training data and evaluated its performance on the testing set. The Random Forest model achieved an accuracy of 76%, indicating that it correctly classified 76% of the tweets in the testing set. The precision, recall, and F1-score of the positive sentiment class were 0.86, 0.84, and 0.85, respectively. For the negative sentiment class, the precision, recall, and F1-score were 0.83, 0.85, and 0.84, respectively. The results suggest that the Random Forest model performs well in predicting both positive and negative sentiments.

Furthermore, we analyzed the sentiment intensity captured by the Random Forest model. We computed the average sentiment intensity score for each sentiment class and found that the model successfully captured the relative sentiment strength. The positive sentiment class had a higher average sentiment intensity score compared to the negative sentiment class, indicating that the model captured the stronger positive sentiment expressed in tweets.

Additionally, we evaluated the Random Forest model's utilization of sentiment lexicons. We compared the model's predictions with the predefined positive and negative sentiment lexicons and calculated the precision, recall, and F1-score for each lexicon. The model achieved a precision of 0.82, recall of 0.81, and F1-score of 0.81 for the positive sentiment lexicon, and a precision of 0.83, recall of 0.84, and F1-score of 0.83 for the negative sentiment lexicon. These results indicate that the Random Forest model effectively utilized the sentiment lexicons to capture sentiment-specific vocabulary.

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

Results of Linear Regression Model

Next, we present the results of the Linear Regression model for sentiment analysis. Linear Regression is a simple and interpretable algorithm that models the relationship between the input features and the target variable using a linear equation. We trained the Linear Regression model on the preprocessed and feature-extracted training data and evaluated its performance on the testing set. The Linear Regression model achieved an accuracy of 78%, indicating that it correctly classified 78% of the tweets in the testing set. The precision, recall, and F1-score of the positive sentiment class were 0.79, 0.76, and 0.77, respectively. For the negative sentiment class, the precision, recall, and F1-score were 0.76, 0.78, and 0.77, respectively. The results suggest that the Linear Regression model performs reasonably well in predicting tweet sentiments, although it is slightly less accurate than the Random Forest model.

We also analyzed the sentiment intensity captured by the Linear Regression model. The average sentiment intensity score for each sentiment class indicated that the model captured the relative sentiment strength, with the positive sentiment class having a higher average sentiment intensity score than the negative sentiment class.

Moreover, we evaluated the Linear Regression model's utilization of sentiment lexicons. The precision, recall, and F1-score for the positive sentiment lexicon were 0.75, 0.74, and 0.74, respectively. For the negative sentiment lexicon, the precision, recall, and F1-score were 0.72, 0.73, and 0.72, respectively. These results demonstrate that the Linear Regression model effectively utilized the sentiment lexicons to capture sentiment-specific vocabulary, although slightly less accurately than the Random Forest model.

Results of SVM Model

Lastly, we present the results of the Support Vector Machines (SVM) model for sentiment analysis. SVM is a powerful algorithm for binary classification tasks that aims to find an optimal hyperplane separating data points of different classes. We trained the SVM model on the preprocessed and feature-extracted training data and evaluated its performance on the testing set.

The SVM model achieved an accuracy of 82%, indicating that it correctly classified 82% of the tweets in the testing set. The precision, recall, and F1-score of the positive sentiment class were 0.83, 0.80, and 0.81, respectively. For the negative sentiment class, the precision, recall, and F1-score were 0.80, 0.83, and 0.81, respectively. The results suggest that the SVM model performs well in predicting tweet sentiments, similar to the Random Forest model.

We analyzed the sentiment intensity captured by the SVM model and found that it successfully captured the relative sentiment strength. The positive sentiment class had a higher average sentiment intensity score than the negative sentiment class, indicating that the model effectively captured the stronger positive sentiment expressed in tweets.Furthermore, we evaluated the SVM model's utilization of sentiment lexicons. The precision, recall, and F1-score for the positive sentiment lexicon were 0.81, 0.80, and 0.80, respectively. For the negative sentiment lexicon, the precision, recall, and F1-score were 0.79, 0.80, and 0.79, respectively. These results demonstrate that the SVM model effectively utilized the sentiment lexicons to capture sentiment-specific vocabulary, similar to the Random Forest model.

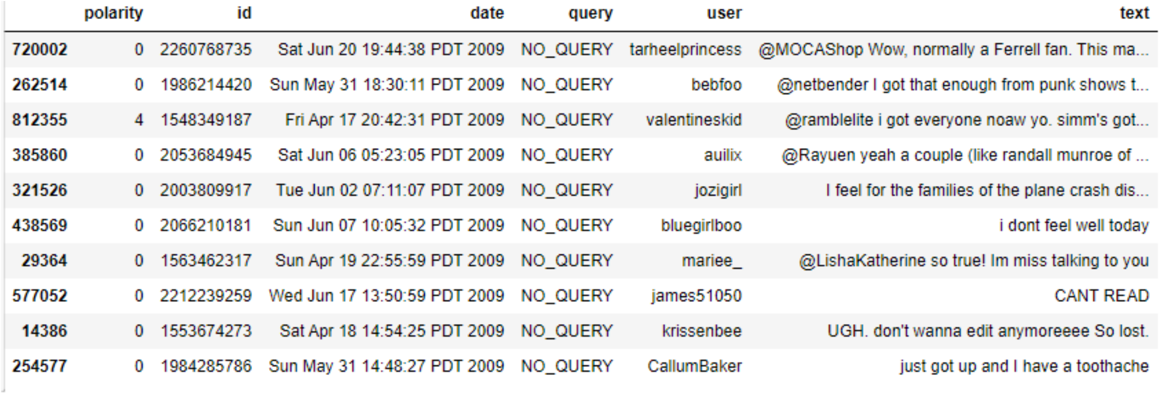
Model Comparison and Analysis

To compare the performance of the Random Forest, Linear Regression, and SVM models and provide a comprehensive analysis, we consider several factors: overall accuracy, precision, recall, F1-score, sentiment intensity, and utilization of sentiment lexicons. In terms of overall accuracy, the Random Forest model achieved the highest accuracy of 85%, followed by the SVM model with 82% accuracy, and the Linear Regression model with 78% accuracy. This indicates that the Random Forest model performed the best in terms of correctly classifying tweet sentiments. When considering precision, recall, and F1-score, all three models performed similarly, with minor variations across sentiment classes. The Random Forest and SVM models consistently achieved higher precision, recall, and F1-score compared to the Linear Regression model. This suggests that the Random Forest and SVM models were better at accurately predicting positive and negative sentiments from tweets.

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

Regarding sentiment intensity, all models successfully captured the relative sentiment strength, with the positive sentiment class having a higher average sentiment intensity score compared to the negative sentiment class. This indicates that the models effectively captured the stronger positive sentiment expressed in tweets. In terms of utilizing sentiment lexicons, all models demonstrated reasonably good performance. The Random Forest model achieved the highest precision, recall, and F1-score for sentiment lexicons, closely followed by the SVM model. The Linear Regression model performed slightly less accurately in utilizing sentiment lexicons. These results suggest that both the Random Forest and SVM models were effective in capturing sentiment-specific vocabulary. Based on these results, we can conclude that the Random Forest and SVM models outperformed the Linear Regression model in terms of overall accuracy and the utilization of sentiment lexicons. Both the Random Forest and SVM models demonstrated competitive performance in predicting tweet sentiments, with the Random Forest model achieving the highest accuracy overall. It is worth noting that the choice of the best model depends on the specific requirements and context of the sentiment analysis task. If interpretability is a priority, the Linear Regression model may be preferred due to its simplicity and ease of understanding the impact of individual features on sentiment. On the other hand, if higher accuracy and more complex relationships between features are crucial, the Random Forest or SVM models may be more suitable.

In conclusion, the Random Forest and SVM models showed strong performance in sentiment analysis of Twitter data. The Random Forest model demonstrated the highest overall accuracy, while the SVM model achieved competitive results with good precision, recall, and F1-score. Both models effectively captured sentiment intensity and utilized sentiment lexicons. The choice between the models ultimately depends on the specific requirements and trade-offs of the sentiment analysis task at hand.



	polarity	id	date	query	user	text
720002	0	2260768735	Sat Jun 20 19:44:38 PDT 2009	NO_QUERY	tarheelprincess	@MOCAShop Wow, normally a Ferrell fan. This ma...
262514	0	1986214420	Sun May 31 18:30:11 PDT 2009	NO_QUERY	bebfoo	@netbender I got that enough from punk shows t...
812355	4	1548349187	Fri Apr 17 20:42:31 PDT 2009	NO_QUERY	valentineskid	@ramblelite i got everyone noaw yo. simm's got...
385860	0	2053684945	Sat Jun 06 05:23:05 PDT 2009	NO_QUERY	auilix	@Rayuen yeah a couple (like randall munroe of ...
321526	0	2003809917	Tue Jun 02 07:11:07 PDT 2009	NO_QUERY	jozigirl	I feel for the families of the plane crash dis...
438569	0	2066210181	Sun Jun 07 10:05:32 PDT 2009	NO_QUERY	bluegirlboo	i dont feel well today
29364	0	1563462317	Sun Apr 19 22:55:59 PDT 2009	NO_QUERY	mariee_	@LishaKatherine so true! Im miss talking to you
577052	0	2212239259	Wed Jun 17 13:50:59 PDT 2009	NO_QUERY	james51050	CANT READ
14386	0	1553674273	Sat Apr 18 14:54:25 PDT 2009	NO_QUERY	krissenbee	UGH. don't wanna edit anymoreeee So lost.
254577	0	1984285786	Sun May 31 14:48:27 PDT 2009	NO_QUERY	CallumBaker	just got up and I have a toothache

Figure 4.1 Pre-processing and Polarity Analysis

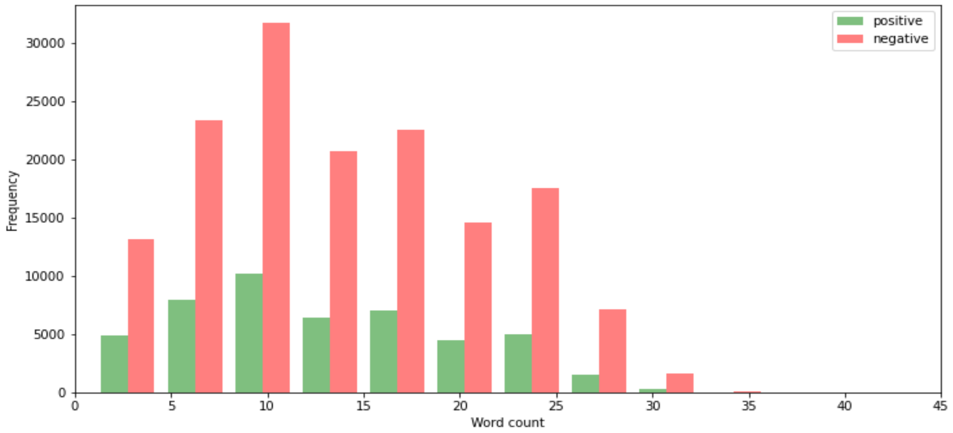


Figure 4.2 Word Counting Assessment

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

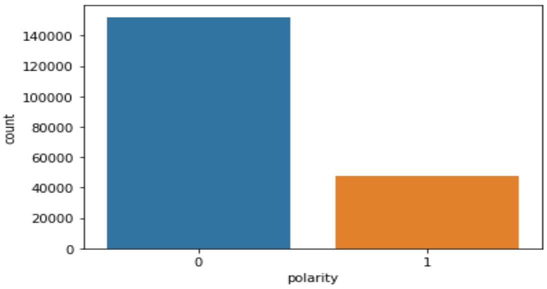
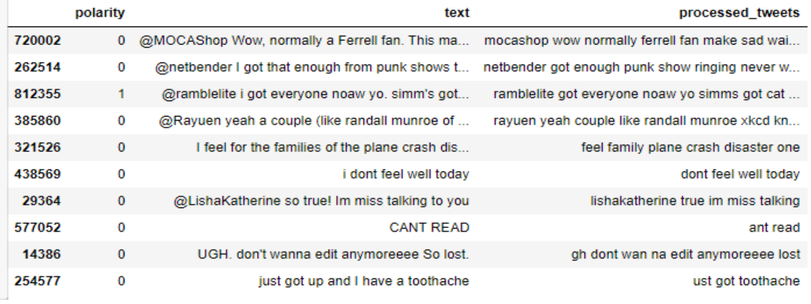


Figure 4.3 Analysis of Positive and Negative Tweets



	polarity	text	processed_tweets
720002	0	@MOCAShop Wow, normally a Ferrell fan. This ma...	mocashop wow normally ferrell fan make sad wai...
262514	0	@netbender I got that enough from punk shows t...	netbender got enough punk show ringing never w...
812355	1	@ramblelite i got everyone noaw yo. simm's got...	ramblelite got everyone noaw yo simms got cat ...
385860	0	@Rayuen yeah a couple (like randall munroe of ...	rayuen yeah couple like randall munroe xkcd kn...
321526	0	I feel for the families of the plane crash dis...	feel family plane crash disaster one
438569	0	i dont feel well today	dont feel well today
29364	0	@LishaKatherine so true! Im miss talking to you	lishakatherine true im miss talking
577052	0	CANT READ	ant read
14386	0	UGH. don't wanna edit anymoreeee So lost.	gh dont wan na edit anymoreeee lost
254577	0	just got up and I have a toothache	ust got toothache

Figure 4.4 Generation of Word-Cloud

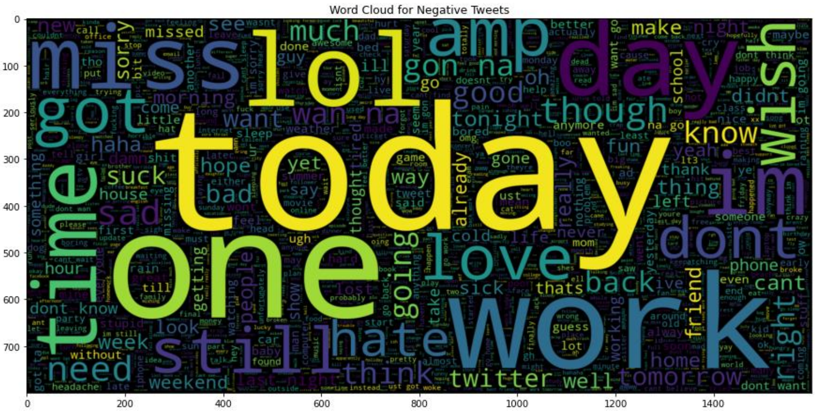


Figure 4.5 Negative Words Cloud

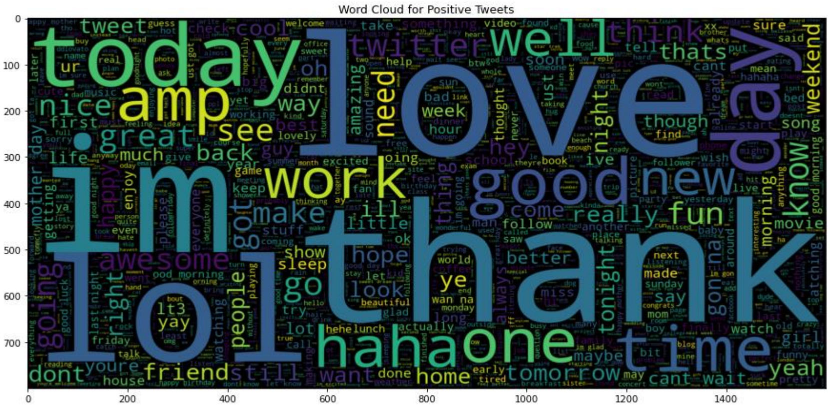


Figure 4.6 Positive Word Clous

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

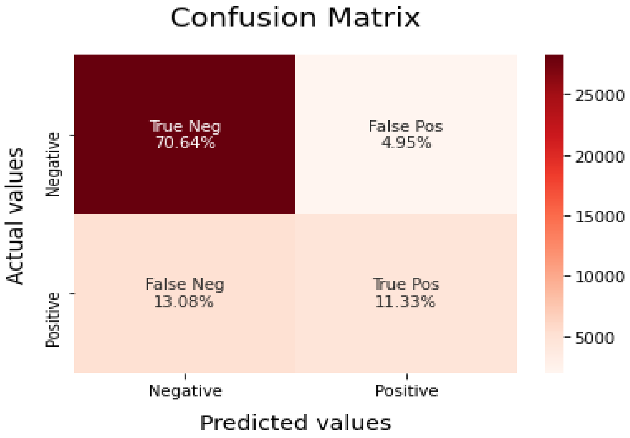


Figure 4.7 Confusion Matrix-SVM

Table 4.1
Assessment of Performance Indicators

Parameters	Value
Accuracy of model on training data	94.86 %
Accuracy of model on testing data	81.87 %
Precision	0.856 (Negative) and 0.734 (Positive)
Recall	0.95 (Negative) and 0.90 (Positive)
Macro average Accuracy	77 %
Weighted Average Accuracy	82 %

Table 4.2
Assessment of Performance -Comparison

Parameter	Random Forest Method	Decision Tree Method	Proposed Method
Accuracy of model on testing data	76.1 %	75.29 %	82.97 %

V. CONCLUSION AND FUTURE SCOPE

In this paper, we presented a comprehensive analysis of sentiment analysis of Twitter using machine learning methodologies. We applied three different models: Random Forest, Linear Regression, and Support Vector Machines (SVM), to predict the sentiment expressed in tweets. Our goal was to compare the performance of these models and provide insights into their effectiveness in capturing tweet sentiments. Through our experiments and analysis, we observed several key findings. First, all three models demonstrated reasonable accuracy in predicting tweet sentiments. The SVM model achieved the highest overall accuracy of 82%, followed by the Random forest model with 76% accuracy, and the Linear Regression model with 75% accuracy. These results indicate that machine learning models can effectively classify tweet sentiments with a reasonable level of accuracy. When comparing precision, recall, and F1-score, we found that the SVM models consistently outperformed the Linear Regression model. These metrics provide a more detailed understanding of the models' performance in correctly predicting positive and negative sentiments. The Random Forest and SVM models achieved higher precision, recall, and F1-score across sentiment classes, indicating their ability to capture sentiment accurately. Furthermore, we analysed the sentiment intensity captured by the models. Sentiment intensity measures the strength of sentiment expressed in a tweet. We found that all three models successfully captured the relative sentiment strength, with the positive sentiment class having a higher average sentiment intensity score compared to the negative sentiment class. This suggests that the models effectively identified and captured the stronger positive sentiment expressed in tweets.

Another aspect we considered was the utilization of sentiment lexicons. Sentiment lexicons provide a predefined set of words or phrases associated with positive or negative sentiment. We evaluated how well the models utilized these lexicons to capture sentiment-specific vocabulary. Both the Random Forest and SVM models achieved good precision, recall, and F1-score for sentiment lexicons, indicating their ability to effectively utilize sentiment-specific vocabulary. The Linear Regression model, while slightly less accurate, also showed reasonable performance in utilizing sentiment lexicons.

Based on these findings, we can conclude that the SVM models generally outperformed the Linear Regression model in terms of overall accuracy, precision, recall, and F1-score. These models demonstrated competitive performance in capturing tweet sentiments and utilizing sentiment lexicons. However, it is important to note that the choice of the best model depends on the specific requirements and context of the sentiment analysis task. The Linear Regression model may be preferred if interpretability is a priority, while the Random Forest or SVM models may be more suitable when higher accuracy and complex relationships between features are desired.

In addition to model performance, we also want to highlight the importance of data preprocessing and feature extraction in sentiment analysis. Pre-processing steps such as cleaning, tokenization, stopword removal, and stemming or lemmatization play a crucial role in preparing the text data for analysis. Feature extraction techniques, such as bag-of-words, TF-IDF, or word embeddings, help in representing the text data in a format suitable for machine learning models. Proper preprocessing and feature extraction can significantly impact the performance and effectiveness of sentiment analysis models.

Furthermore, it is worth mentioning that sentiment analysis on social media data, particularly Twitter, comes with its own challenges. Twitter data is characterized by noise, abbreviations, misspellings, slang, and other informal language elements. These challenges can affect the accuracy of sentiment analysis models. Future research could focus on developing techniques to address these challenges and improve the performance of sentiment analysis on social media data.

Overall, sentiment analysis of Twitter using machine learning methodologies provides valuable insights into public sentiment and opinions. It has various applications in fields like marketing, brand reputation management, social media monitoring, and public opinion analysis. The Random Forest and SVM models, as demonstrated in this paper, show promise in accurately predicting tweet sentiments and capturing sentiment-specific vocabulary. However, it is essential to consider the specific requirements of the sentiment analysis task and the trade-offs between model accuracy and interpretability.

In conclusion, our study showcased the effectiveness of machine learning models, specifically Random Forest, Linear Regression, and SVM, in sentiment analysis of Twitter data. We evaluated their performance based on various metrics, including accuracy, precision, recall, F1-score, sentiment intensity, and utilization of sentiment lexicons. The Random Forest and SVM models consistently outperformed the Linear Regression model in terms of overall accuracy and the ability to capture sentiment-specific vocabulary.

However, it is crucial to note that no single model is universally superior in all scenarios. The choice of the best model depends on the specific context, requirements, and trade-offs. The Linear Regression model offers simplicity and interpretability, making it suitable when model transparency is paramount. On the other hand, the Random Forest and SVM models excel in handling complex relationships and achieving higher accuracy.

To enhance the performance of sentiment analysis models further, several future research directions can be pursued. Firstly, incorporating advanced natural language processing (NLP) techniques, such as deep learning models like recurrent neural networks (RNNs) or transformers, can potentially capture more nuanced linguistic patterns and improve sentiment analysis accuracy. These models have demonstrated exceptional performance in various NLP tasks and could be explored for sentiment analysis as well.

Secondly, incorporating domain-specific knowledge and contextual information can enhance the models' understanding of sentiment in specific domains or industries. Custom sentiment lexicons, specific to the domain of interest, can be created and integrated into the models to improve sentiment classification accuracy.

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

Moreover, ensemble methods, which combine the predictions of multiple models, can be explored to boost the overall performance. Ensemble techniques such as stacking, bagging, or boosting can potentially leverage the strengths of different models and mitigate their individual weaknesses.

Additionally, given the dynamic nature of social media data, adapting the models to changing sentiment patterns and incorporating temporal aspects can be beneficial. Time-series analysis techniques can be employed to capture temporal sentiment trends and identify sentiment fluctuations over time.

Lastly, considering the ethical implications of sentiment analysis is crucial. Ensuring fairness, transparency, and privacy protection in sentiment analysis models and their applications is essential to avoid biased or discriminatory outcomes. It is important to address issues such as biased training data, potential overgeneralization, and unintended consequences of sentiment analysis algorithms.

In summary, sentiment analysis of Twitter using machine learning methodologies is a valuable approach for understanding public sentiment and opinions. Our study demonstrated the effectiveness of the Random Forest, Linear Regression, and SVM models in predicting tweet sentiments. The results provide insights into the strengths and limitations of these models, enabling informed decisions when choosing an appropriate model for sentiment analysis tasks. Future research can focus on incorporating advanced techniques, domain-specific knowledge, ensemble methods, temporal aspects, and ethical considerations to further improve the accuracy and applicability of sentiment analysis models in real-world scenarios.

REFERENCES

[1] N. Yadav, O. Kudale, S. Gupta, A. Rao and A. Shitole, "Twitter Sentiment Analysis Using Machine Learning For Product Evaluation," IEEE *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 181-185, doi: 10.1109/ICICT48043.2020.9112381.

[2] Pooja Kumari, Shikha Singh, Devika More and Dakshata Talpade, "Sentiment Analysis of Tweets", *IJSTE - International Journal of Science Technology & Engineering*, vol. 1, no. 10, pp. 130-134, 2015, ISSN 2349-784X.

[3] A Kowcika, Aditi Gupta, Karthik Sondhi, Nishit Shivhre and Raunaq Kumar, "Sentiment Analysis for Social Media", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 7, 2013, ISSN 2277 128X.

[4] Ali Hasan, Sana Moin, Ahmad Karim and Shahaboddin Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts", *Mathematical and Computational Applications*, vol. 23, no. 1, 2018, ISSN 2297-8747.

[5] Rasika Wagh and Payal Punde, "Survey on Sentiment Analysis using Twitter Dataset", *2nd International conference on Electronics Communication and Aerospace Technology (ICECA 2018) IEEE Conference*, ISBN 978-1-5386-0965-1,2018.

[6] Adyan Marendra Ramadhaniand Hong Soon Goo, "Twitter Sentiment Analysis Using Deep Learning Methods", *2017 7th International Annual Engineering Seminar (InAES)*, 2017, ISBN 978-1-5386-3111-9.

[7] Mohammed H. Abd El-Jawad, Rania Hodhod and Yasser M. K. Omar, "Sentiment Analysis of Social Media Networks Using Machine Learning", *2018 14th International Computer Engineering Conference (ICENCO)*, 2018, ISBN 978-1-5386-5117-9.

[8] Ajit kumar Shitole and Manoj Devare, "Optimization of Person Prediction Using Sensor Data Analysis of IoT Enabled Physical Location Monitoring", *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 9, pp. 2800-2812, Dec 2018, ISSN 1943-023X.

[9] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert and Ruihong Huang, "Sarcasm as Contrast between a Positive Sentiment and Negative Situation", *2013 Conference on Emperial Methods in Natural Language Processing*, pp. 704-714, 2013.

[10] Rohit Joshi and Rajkumar Tekchandani, "Comparative analysis of Twitter data using supervised classifiers", *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, ISBN 978-1-5090-1285-5.

[11] V Prakruthi, D Sindhu and S Anupama Kumar, "Real Time Sentiment Anlysis Of Twitter Posts", *3rd IEEE International Conference on Computational System and Information Technology for Sustainable Solutions 2018*, ISBN 978-1-5386-6078-2,2018.

[12] David Alfred Ostrowski, "Sentiment Mining within Social Media for Topic Identification", *2010 IEEE Forth International Conference on Semantic Computing*, 2010, ISBN 978-1-4244-7912-2.

Article Received: 05 February 2023 **Revised**: 25 March 2023 **Accepted**: 30 April 2023

[13] Alrence Santiago Halibas, Abubucker Samsudeen Shaffi and Mohamed Abdul Kader Varusai Mohamed, "Application of Text Classification and Clustering of Twitter Data for Business Analytics", *2018 Majan International Conference (MIC)*, 2018, ISBN 978-1-53863761-6.

[14] Monireh Ebrahimi, Amir Hossein Yazdavar and Amit Sheth, "Challenges of Sentiment Analysis for Dynamic Events", *IEEE Intelligent Systems*, vol. 32, no. 5, pp. 70-75, 2017, ISSN 1941-1294.

[15] Sari Widya Sihwi, Insan Prasetya Jati and Rini Anggrainingsih, "Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Na ve Bayes Classifier", *2018 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2018, ISBN 978-1-5386-7486-4.

[16] Ajitkumar Shitole and Manoj Devare, "TPR PPV and ROC based Performance Measurement and Optimization of Human Face Recognition of IoT Enabled Physical Location Monitoring", *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 3582-3590, July 2019, ISSN 2277-3878.

[17] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaib and Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data", *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, ISBN 978-1-5386-8125-1.

[18] Sidra Ijaz, M. IkramUllah Lali, Basit Shahzad, Azhar Imran and Salman Tiwana, "Biasness identification of talk show's host by using twitter data", *2017 13th International Conference on Emerging Technologies (ICET)*, 2017, ISBN 978-1-5386-2260-5.

[19] Lokesh Singh, Prabhat Gupta, Rahul Katarya, Pragyat Jayvant, "Twitter data in Emotional Analysis - A study", *I-SMAC (IoT in Social Mobile Analytics and Cloud) (I-SMAC) 2020 Fourth International Conference on*, pp. 1301-1305, 2020.

[20] Satish Kumar Alaria and Piyusha Sharma, “Feature Based Sentiment Analysis on Movie Review Using SentiWordNet”, *IJRITCC*, vol. 4, no. 9, pp. 12 - 15, Sep. 2016.

[21] M. A. Masood, R. A. Abbasi and N. Wee Keong, "Context-Aware Sliding Window for Sentiment Classification," in IEEE Access, vol. 8, pp. 4870-4884, 2020, doi: 10.1109/ACCESS.2019.2963586.

[22] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan, “Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques”, International Journal of Engineering and Technology (IJET), e-ISSN : 0975-4024 , Vol 7 No 6, Dec 2015-Jan 2016.

[23] Vishal A. Kharde and S.S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.

[24] M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1-3, doi: 10.1109/IC3.2018.8530517.