Journal of Multi Disciplinary Engineering Technologies
Volume 18, Issue 02

Emotions in Tweets: A Sentiment Analysis Approach

Diksha Rawat[1], Manik Arora[1], Yash Khandelwal[1], Shahzeb Khan[2*]

Department of Computer Science Application, Sharda University, Greater Noida, U.P., 201310, India.
[2]Assistant Professor, Department of Computer Science Application, Sharda University, Greater Noida, U.P., 201310, India.

*Corresponding author(s). E-mail(s): shahzeb.khan@sharda.ac.in;
Contributing authors: dikshu9705@gmail.com; aroramanik539@gmail.com;
yashkhandelwal2604@gmail.com;

Abstract

Social media sentiment research has evolved into a crucial tool for analysing public opinion, brand perception, and customer sentiment in the digital age. This approach uses natural language processing (NLP), machine learning, and computational linguistics to categorize the attitudes conveyed in social media posts as positive, negative, neutral, or irrelevant. Because social media writing is unstructured and informal, the work provides distinct difficulties, such as dealing with slang, sarcasm, and multilingual content. The technique for developing a sentiment analysis system often begins with data collecting from GitHub or Twitter (using a separate account). To prepare for analysis, the data is pre processed using tokenization, Potter stemming, Text Blob, word tokenization, stop words, word clouds, and text normalization. Feature extraction approaches, such as Count Vectorizer, turn the text into numerical representations, which are then used to train machine learning models like Logistic Regression and SVM, were trained and assessed using F1- score, recall, accuracy, and precision. With the highest accuracy of 97.49%, the SVM model demonstrated its superiority over Logistic Regression in sentiment categorization.

Keywords: Sentiment Analysis, Twitter Data, Natural Language Processing (NLP), Machine Learning, Social Media Analytics, Logistic Regression, Support Vector Machine (SVM), Data Preprocessing, Polarity Detection, Feature Extraction

1 Introduction

n today's digital landscape, extensive text data from social media, reviews, and forums provides valuable insights into sentiment and public opinion. Sentiment analysis is valuable for addressing a variety of challenges relevant to humancomputer interaction experts, researchers, and professionals in fields like sociology, marketing, advertising, psychology, economics, and political science. [1] In recent years, these technologies have been expanded to examine other aspects, such as a user's stance on a specific topic. Living in society, people form opinions—both positive and negative—about

Published on: May, 2025 ISSN(P) 0974-1771 ISSN(O) 2581-9372

individuals, products, places, and events. These opinions can be categorized as sentiments [2]. Sentiment analysis involves the development of automated methods for identifying and extracting sentiments from written text. [3]. Sentiment analysis, a branch of natural language processing, focuses on interpreting and quantifying emotions expressed in text. Also referred to as opinion mining, it is a research domain aimed at examining people's sentiments or opinions about various entities, including topics, events, individuals, issues, services, products, organizations, and their characteristics [4]. The sentiment found within comments, feedback or critiques provide useful in dicators form any different purposes [5].

Sentiment analysis facilitate s the understanding of trends and attitudes by corporations, governments, and researchers by differentiating between sentiments such as positive, negative, and neutral. The field of Sentiment Analysis aims to understand these opinions and classify them into categories such as positive, negative, or neutral. So far, the majority of sentiment analysis research has focused on review platforms [6]. The intricacies of human emotion, such sarcasm and context, are still difficult to reliably capture in machine learning and deep learning models, despite tremendous advances in these fields. Before the advent of the Web, individuals typically sought opinions from friends and family when making decisions, while organizations relied on methods like opinion polls, surveys, and focus groups to gauge public sentiment about their products and services. [7] However, the rapid growth of user-generated content on the Web in recent years has revolutionized this process. Consequently, a sentiment analysis system designed to handle large datasets (such as those on platforms like Twitter) must be capable of processing text efficiently and at high speed [8]. This study examines data collection and pre-processing, utilizing visual aids such as word clouds, bar graphs, pie charts, and confusion matrices to illustrate our data sentiments.

It centers on determining the appropriate level of accuracy through a comparison of two models (logistic regression and SVM), as well as checks on precision, recall, and F1 score. The findings aim to increase accuracy by using a customized model and sentiment analysis from a variety of fields. We will talk about the dataset and how the model uses it in this research paper. We will also examine which classification is more accurate than the others and why. The study aims to enhance accuracy by applying sentiment analysis across multiple domains and creating a customized model. In this research article, we analyses the dataset and model to determine which categorization has more accuracy and why. The findings aim to increase accuracy by combining sentiment analysis across many domains with a personalized model. Our dataset contains around 75682 sentiments and tweets. The data information we looked at from our entire dataset (75682) included non-null values for sentiment (75682), social media (75682), and tweets (74996). Tweets display 686 null values in our text df.isnull. We will begin by pre-processing our data, which entails using a tokenizer to split unstructured text data into smaller units called tokens. A word, a single character, or considerably larger textual entities can all be considered tokens. Porter stemming will be used to eliminate English word suffixes in the following phase. When talking about information retrieval, the capability to automatically remove suffixes is really helpful. Count Vectorizer, which converts text into numerical representations, will be used for our feature extraction. In order to determine which classification has higher accuracy than the others and why, we will examine the dataset and how it functions in the model in this research article.

2 LITRATURE SURVEY

The research paper presents VADER, a rule-based sentiment analysis model for tweets and other microblogs. VADER distinguishes between sentiment intensity and tackles social media issues such as emoticons, slang, and acronyms. It exhibits excellent generalizability across a variety of text contexts, such as movie and product reviews, and is computationally efficient and domain-agnostic. [1] The study examines the most recent advancements in social media sentiment analysis, emphasizing novel techniques such as sentiment-aware word embeddings, convolutional neural networks, and multilevel semantic network visualization. Along with examining applications in cyber-aggression

detection, e-commerce psychographic segmentation, gender detection, and health insurance attitude tracking, it also highlights the development of the field through the use of big data and deep learning [2]. The study addresses language ambiguity and context dependent terms in its discussion of sentiment analysis on social media. It examines current tools and approaches and suggests a solution for word meaning disambiguation and polarity scoring that makes use of SentiWordNet and supervised learning techniques. Product profiling, trend analysis, and trend visualization are all done by the system using data from Twitter. [3] This study explores sentiment analysis techniques in social media environments, categorizing research into methodology oriented, granularity-oriented, and task-oriented approaches. It addresses challenges like negation handling, implicit sentiment identification, and cross-domain analysis. The paper provides a methodical resource for both novices and experts, discussing tools, datasets, and future approaches like multimodal sentiment analysis [4]. The study explores sentiment analysis techniques using a hybrid methodology combining machine learning, supervised, and rule-based methods. It covers hybrid classification models, machine learning algorithms, and pattern-based natural language processing. The review covers tasks like sentiment categorization, prediction at granular levels, and aspect-based analysis. It also addresses challenges like feature selection, sentiment ambiguity, and data sparsity in automated classification systems [5]. The study highlights Twitter's significance as a source of public opinion by examining sentiment analysis techniques for its data. It tackles issues brought on by tweets' casual style and conciseness. To increase the accuracy of sentiment categorization, the authors suggest a hybrid strategy that combines dictionary-based and corpus-based techniques with machine learning and natural language processing. A case study illustrates the efficacy of the method [5]. The study offers a thorough examination of sentiment analysis with an emphasis on gleaning subjective view points from written material. Comparative sentence processing, feature based analysis, and sentiment categorization at the document and sentence levels are all covered. The study employs both supervised and unsupervised methodologies, tackles issues such as implicit features, domain dependency, and opinion spam, and makes use of visualizations for real-world usage in decision-making, product development, and marketing [7]. The study offers a hybrid approach to sentiment analysis on Twitter that combines sentiment lexicons, minimum linguistic processing, and supervised learning. It tackles issues including brief text, colloquial language, and the need for real-time processing. By normalizing slang, emoticons, and punctuation, the technique enhances sentiment classification performance across a variety of datasets, including blog posts and tweets. This method provides information for sentiment analysis applications in many languages and real-time [8].

The study examines machine learning and lexicon-based approaches to sentiment analysis in social media. It points out that Facebook struggles with unstructured and noisy data, while Twitter is the most widely used site for sentiment data extraction because of its structured format and API accessibility. Applications for sentiment analysis are numerous and include politics, business, healthcare, disaster relief, and social trends. In order to improve sentiment analysis capabilities, the study highlights the need to combine approaches and investigating neglected data sources [9]. The study explores the use of sentiment analysis (SA) in formative evaluation in higher education, focusing on online and hybrid learning settings. It highlights the growing use of SA for forecasting learning outcomes, improving feedback, and assessing student emotions. However, the study finds that current SA tools lack gender-sensitive and culturally inclusive approaches. The authors aim to address research and practice gaps for more inclusive educational assessment [10]. The study examines sentiment analysis (SA) in social networks, emphasizing its uses in emergency management, politics, economics, marketing, and health. Given that Twitter is a major data source, it emphasizes the necessity of both conventional and cutting-edge approaches. The study emphasizes the need for more research in understudied areas by highlighting gaps in repeatability, AI integration, and real-world applicability [11].

In order to increase accuracy, especially in the classification of negative sentiment, the study investigates the use of semantic characteristics in sentiment classifiers. It assesses three techniques: interpolation, augmentation, and replacement. The findings indicate that while sentiment-topic

features outperform semantic features in topic-specific datasets, semantic features improve accuracy in generic datasets. Refinement of entity specificity and application of the method to wider contexts are future directions [12].

The paper discusses the importance of sentiment analysis in social media, focusing on Twitter as a microblogging platform. It discusses machine learning algorithms used for sentiment classification, including Naive Bayes, MaxEnt, and SVM. The study also highlights the challenges of sentiment analysis due to informal tweets with emoticons and acronyms. Key contributions include Parts Of Speech specific prior polarity features and a tree kernel approach. The authors propose a hybrid approach combining corpus-based and dictionary-based techniques for improved sentiment detection [13]. The study provides a comprehensive overview of sentiment analysis, highlighting its techniques, uses, and challenges. It categorizes techniques into lexicon-based, machine learning based, and hybrid approaches. It highlights its applications in social media monitoring, corporate intelligence, and healthcare, highlighting its usefulness in understanding customer behavior and maintaining brand reputation. It also addresses issues like high-quality training data and sarcasm detection [14].

The study examines sentiment analysis techniques for social media platforms, assessing machine learning classifiers like SVM, Naive Bayes, Random Forest, and Linear Regression in addition to feature extraction techniques like Part-of-Speech tagging (POS), Bag-of-Words (BoW), and hashtag analysis. It also draws attention to issues such as managing sarcasm, symbols, and subtleties in language, as well as the dearth of annotated training datasets. According to the study, future research should concentrate on enhancing sentiment analysis that is multilingual, multimodal, and context-aware [15] This study compares various sentiment analysis models on Twitter, including a unigram model and tree kernel techniques. Results show feature-based and tree kernel models outperform the unigram baseline in both three-way and binary tasks, with a significant improvement of over 4%. The study suggests further research into in-depth language studies and concludes that sentiment analysis for Twitter data is comparable to other genres [16].

The study examines sentiment analysis approaches and techniques for Twitter data, with an emphasis on the microb logging platform's characteristics and drawbacks, such as its character restriction. It talks about supervised and unsupervised learning techniques and divides sentiment analysis into document, sentence, and entity/aspect levels. The study examines several machine learning methods and outlines how they might be used in public opinion tracking, marketing, and finance [17]. The study explores sentiment analysis, focusing on identifying positive and negative emotions in various contexts like political forecasting and product reputation evaluation. It compares sentiment analysis with other analytical fields and highlights the challenges of skewed data, particularly in social media. The study also highlights the ongoing debate on the usefulness of sentiment analysis methods and their practical applications [18].

The study examines sentiment analysis on Twitter data, emphasizing its value in disaster relief, political polling, and marketing. In addition to comparing supervised machine learning techniques like Naıve Bayes, Support Vector Machines, and Random Forests, it talks about how feelings are divided into positive, negative, and neutral categories. Future research directions are also suggested in the report, such as tackling language diversity in social media data and using domainagnostic sentiment categorization systems [19].

The study uses Twitter data to analyze public opinion on Indonesia's 2019 presidential candidates, using the Naıve Bayes algorithm for sentiment categorization. The algorithm outperforms SVM and KNN with an accuracy of 80.90The study highlights the importance of web crawlers and sentiment analysis in understanding public opinion. It compares classification techniques and discusses data processing methodologies [20].

The research paper discusses sentiment analysis, its importance in academic and business contexts, and its evolution from tracking public opinion to understanding sentiment variations for decision-making. It reviews various methodologies, including Natural Language Processing (NLP), statistics, and machine learning techniques like Na¨ıve Bayes, Support Vector Machines (SVM), and Latent Dirichlet Allocation (LDA). The authors highlight the relevance of sentiment analysis in

understanding consumer behavior and suggest future research directions to improve its application across various domains [21].

The study offers a thorough examination of Twitter Sentiment Analysis (TSA) techniques, emphasizing the use of cognitive science ideas, machine learning, and semantic technologies. It draws attention to the difficulties posed by big data and the necessity of cutting-edge tools like Hadoop and Apache Spark. The study classifies current TSA practices, such as SNA measurements and visualization strategies, and comes to the conclusion that TSA is a dynamic field with room to grow [22].

The study examines sentiment analysis techniques with an emphasis on Twitter data and their use in gauging public opinion on a range of subjects. It demonstrates the efficiency of machine learning classifiers like SVM and Naive Bayes and divides methods into document-level, sentence-level, and aspect-level classifications. The report discusses issues and makes recommendations for future lines of inquiry [23]. This study explores sentiment analysis methods using machine learning strategies, including neural networks, SVM, and Naive Bayes, on Twitter data. It discusses challenges like dealing with noisy data and polarity shifts and emphasizes the importance of feature extraction, ensemble approaches, nd preprocessing techniques for model accuracy. The authors conclude that while machine learning has shown promising results, further research is needed to improve sentiment analysis models' functionality [24].

An important component of natural language processing, phrase-level sentiment analysis, is examined in this research study along with its relationship to context. It draws attention to the difficulties in interpreting sentiment in sentences, which can differ greatly. By taking contextual elements into account, the authors suggest ways to increase the accuracy of sentiment recognition and improve comprehension of user sentiment across a range of applications [25].

3 DATA DESCRIPTION

3.1 Exploratory Data Analysis (EDA)

The initial stage in conducting sentiment analysis on Twitter is gathering pertinent information from the network. We have gathered 75,682 pieces of data for this using GitHub. If you want to gather data via an API, you must do the following: Keywords and Hashtags: Locating and choosing pertinent hashtags and keywords associated with the topic of interest. This facilitates the collection of tweets that address particular topics or occasions. Time Frame: Establishing the duration of the data collection process to guarantee the dataset's manageability and relevancy Filtering Criteria: Using filters to make sure the dataset reflects a range of viewpoints and to remove irrelevant tweets (such as spam).

3.2 Data Preprocessing

To improve the quality of the data for sentiment analysis, raw tweets are pre-processed. This includes:

Text Cleaning: Removing unnecessary elements such as URLs, special characters, and emojis that do not contribute to sentiment analysis.

Tokenization: Splitting tweets into individual tokens or words.

Stop Word Removal: Eliminating common words that do not carry significant meaning in sentiment analysis (e.g., "and," "the").

Stemming/Lemmatization: Reducing words to their base or root form to standardize the dataset (e.g., "running" to "run").

Following the use of Potter Stemming, we looked at the non-null values tweets (68926) among the data from our entire dataset (75682). Tweets display 1 null value in our text df.isnull.

Polarity: The sentiment of a text is indicated by its polarity, which is usually classified as neutral, negative, or positive. Opinions in reviews, social media, etc. are analysed using it.

	Social Media	Sentiments	Tweets
0	Borderlands	Positive	im getting on borderlands and i will murder yo...
1	Borderlands	Positive	I am coming to the borders and I will kill you...
2	Borderlands	Positive	im getting on borderlands and i will kill you ...
3	Borderlands	Positive	im coming on borderlands and i will murder you...
4	Borderlands	Positive	im getting on borderlands 2 and i will murder ...

Fig. 1: Data

In our model the polarity scores are assigned as Value ¿ 0 = positive, 0 = neutral, Value ¡ 0 = Negative. For Positive, we have total 26703 rows as positive showing value ¿ 0, for Negative, we have total 19711 rows as negative showing value ¡ 0, for Neutral, we have total 22513 rows as neutral showing value = 0.

Since 6395 was our irrelevant sentiment in the dataset, our total dataset of 75682 was adjusted to 69287. We re-examined the data information from our entire dataset (69287) after completing all of the data preprocessing. This time, Sentiment (68927), Polarity (68927), and Tweets (69287) all had non-null values and no null values remained.

4 METHODOLOGY

The core of the methodology involves applying sentiment analysis techniques to the pre processed tweets. This process includes:

4.1 Model Selection

Selecting suitable models for sentiment analysis. In this we have used Logistic Regression and SVM (Support Vector Machine). We have chosen these with the help of eager learning (Builds a model during training, which allows for quicker predictions at query time by using the pre-built model).

4.2 Training and Testing

4.2.1 Splitting the Dataset

- The line in our code "x train, x test, y train, y test = train test split(X, Y, test size = 0.2, random state = 42)" suggests that the dataset is divided into training and testing sets.
- Explanation: X and Y represent features and labels, respectively. Here, X could be the text data (tweets), and Y could be the sentiment labels.
- Train test split splits X and Y into x train, y train (for training) and x test, y test (for testing).
- Test size = 0.2 means that 20for testing, while 80
- Random state ensures reproducibility by controlling the random split

4.2.2 Model Training

- You might see something like model.fit(x train, y train), which fits a machine learning model (e.g., LogisticRegression) on the training data.
- This step enables the model to learn patterns in the data so it can make predictions on unseen data.

4.2.3 Model Testing/Evaluation

- In the evaluation step, the model's performance is tested using the test set.
- Metrics like accuracy score, classification report, and confusion matrix from sklearn measure how well the model performs.

- Example usage might be accuracy score(y test, model.predict(x test)), which compares predicted labels with actual labels in the test set.

4.2.4 Sentiment Classification

Applying the trained models to classify tweets into sentiment categories (e.g., positive, negative, neutral). We have classified with the help of polarity scores as +1, 0, and -1.

4.3 Evaluation

To assess the performance of the sentiment analysis models, the following evaluation metrics are used:

Accuracy: The proportion of correctly classified tweets out of the total number of tweets.

Precision, Recall, and F1-Score: These measures shed light on how well the model distinguishes between positive, negative, and neutral moods.

Confusion Matrix: Analyzing the confusion matrix helps in understanding the distribution of true positives, true negatives, false positives, and false negatives.

Word Cloud: A word cloud highlighting frequently mentioned terms associated with different sentiment categories.

4.4 Analysis and Interpretation

The classified sentiments are analyzed to draw insights and conclusions. This includes:

Sentiment Distribution: Examining the proportion of each sentiment category to understand overall public opinion.

Comparative Analysis: Comparing results across different keywords, hashtags, or time periods to identify significant variations or trends

4.5 Visualizations

Graphs and charts to visually represent sentiment trends and distributions.

4.5.1 Bar Plot

Rectangular bars are used to graphically display data in a bar plot, also called a bar chart. Each bar has a length that corresponds to the value it represents; the bars are usually positioned either horizontally or vertically.

- Neutral: With a total just above 20,000, this bar shows that more than 20,000 tweets were categorized as having a neutral sentiment.
- Positive: With a count of about 27,000, this is the highest bar and indicates that the most prevalent sentiment in the examined tweets was positive.
- Negative: With almost 20,000 tweets expressing negative sentiment, this bar is just below the neutral count.

All things considered, the most common sentiment in the examined tweets was positive, which was followed by neutral and finally negative. This implies that the tweets gathered for this investigation had a generally upbeat attitude.

4.5.2 Pie Chart

A pie chart illustrating the proportion of positive, negative, and neutral sentiment.

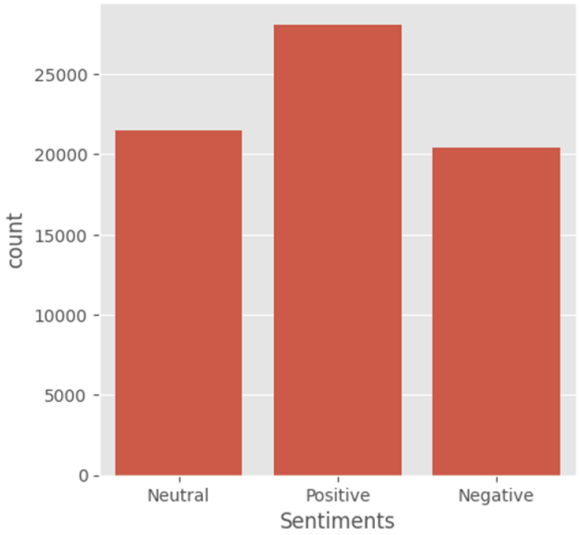

Fig. 2: Bar Graph

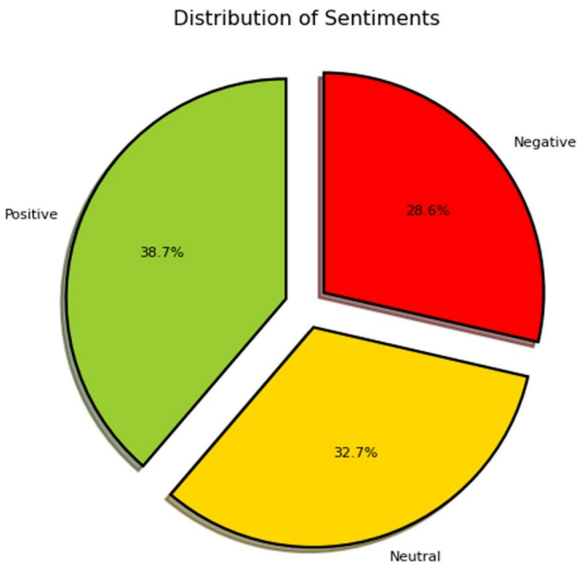

Fig. 3: Pie Chart

4.5.3 Scatter plot

One kind of graph used to show the relationship between two variables is a scatter plot. Two coordinates—one on the x- and one on the y-axes—determine the position of each point on the plot, which represents a data point.

4.5.4 Diagonal plots

These display each sentiment component's distribution (density plots), such as vader neg and roberta pos.

- It is simpler to observe how scores are allocated for each sentiment type since each sentiment label—Positive, Neutral, Negative, and Irrelevant—is color-coded (Blue, Orange, Green, and Red,

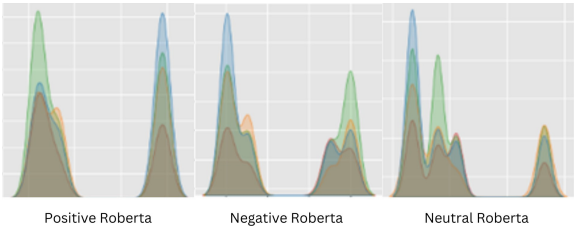

Fig. 4: Roberta

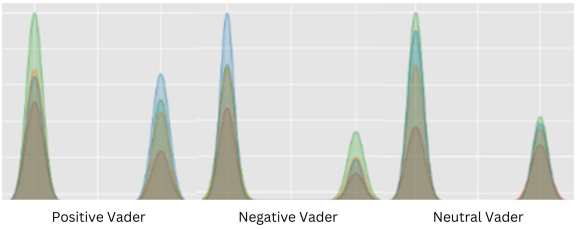

Fig. 5: Vander

respectively). For instance, vader pos has a peak near zero for irrelevant and negative feelings, and larger scores for positive sentiments.

- Plots that are off-diagonal: These scatter plots demonstrate the connections between various sentiment scores. Vader neg vs. roberta neg, for instance, illustrates the differences between the negative sentiment scores from RoBERTa and VADER. The majority of data points do not appear to clearly reflect negative emotion, according to points clustered closer to zero. As you can see, points with neutral or positive sentiment (based on colour) tend to group together in particular regions, indicating that the models' sentiment detection is consistent.
- Clusters of Sentiment: In both models, negative feelings (orange) tend to emerge in clusters with greater neg values, whereas positive sentiments (blue) typically cluster at higher pos values. For both models, neutral attitudes (green) tend to cluster in the middle of the score distributions.

4.5.5 Confusion matrix

A confusion matrix is used in sentiment analysis to assess how successfully a model categorizes text data into various sentiment categories (e.g., positive, negative, or neutral). By contrasting the actual and predicted sentiment labels, it is possible to evaluate how accurate the sentiment predictions were.

Diagonal Cells: These show accurate forecasts in which the true label and the expected label coincide:

- Negative-Negative: 3719 tweets were accurately predicted to be negative (negative-negative).
- Neutral-Neutral (4443): The prediction of neutral was accurate for 4443 tweets.
- Positive-Positive: 5176 tweets were accurately classified as positive (positive-positive).

The model accurately predicted a significant number of tweets in each sentiment category, as indicated by the values in these cells, which are the highest in each row. Misclassifications are represented by off-diagonal cells, in which the true label differs from the anticipated label:

- Negative-Neutral (82): 82 tweets were predicted to be neutral even though they were actually negative.
- Negative-Positive (105): 105 tweets were projected to be positive when they were actually negative.

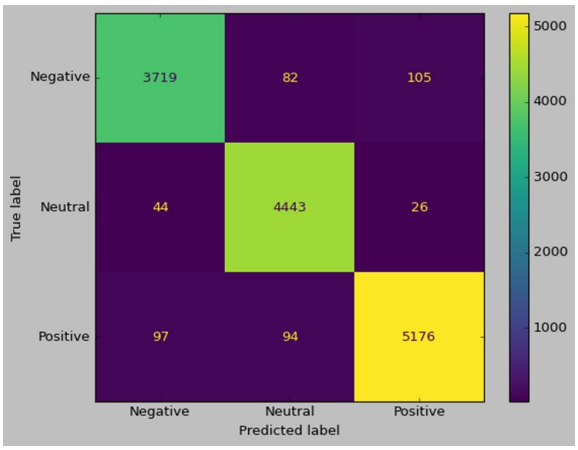

Fig. 6: Confusion Matrix

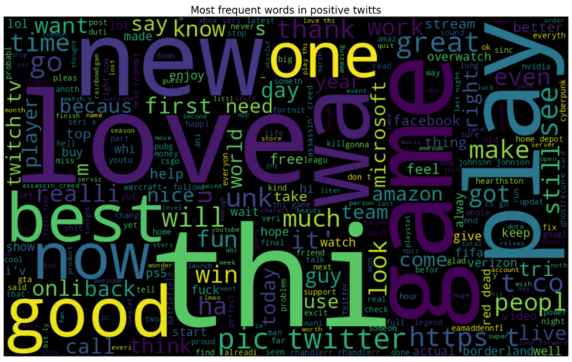

Fig. 7: Positive Word Cloud

- Neutral-Negative (44): 44 tweets were projected to be negative even though they were actually neutral.
- Neutral-Positive (26): 26 tweets were projected to be positive even though they were actually neutral.
- Positive-Negative (97): 97 tweets were predicted to be negative even though they were actually positive.
- Positive-Neutral(94): 94 tweets that were truly positive were forecasted to be neutral, resulting in a Positive- Neutral score.

4.5.6 Word Cloud

A word cloud is a visual representation of text data where the size of each word reflects its frequency or significance. Less frequent words are displayed in smaller fonts, while words that occur more frequently in the dataset are displayed in larger, bolder fonts. Word clouds are frequently used to rapidly pinpoint the most important themes or subjects within a sizable body of text, such as comments on social media, reviews, or survey answers.

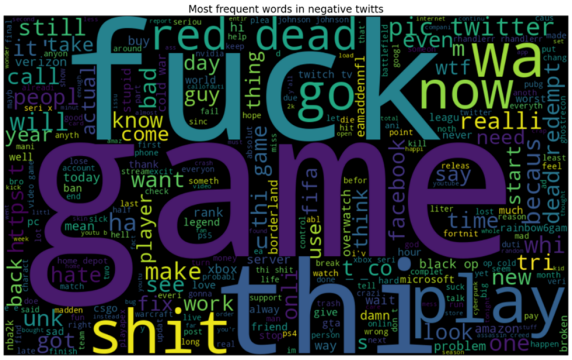

Fig. 8: Negative Word Cloud

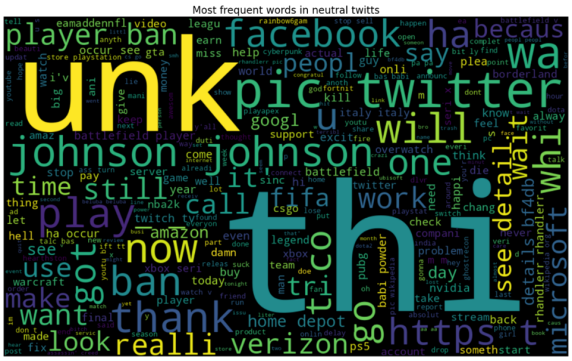

Fig. 9: Neutral Word Cloud

4.5.7 Heatmap

A heatmap is a visual representation of data where colors are used to indicate the values of individual elements within a matrix. It is widely used in data analysis to display the intensity or density of values across multiple dimensions, making it simple to discover trends, correlations, and outliers.

High Correlation: Cells with values close to 1 (e.g., 0.94 at (1, 1)) indicate a strong positive relationship between the features. For sentiment analysis, a high correlation could mean that two sentiment indicators often occur together or have a similar distribution. **Low or Negative Correlation:** Cells with values close to 0 or negative (e.g., 0.036 at (0, 4)) indicate little to no relationship. In sentiment analysis, this might suggest that the two features are independent, meaning one does not influence the other. **Diagonal Values:** The diagonal cells (top left to bottom right) have a value of 1.0 (or close to it), as they represent the correlation of each feature with itself.

4.6 Future Use

Sentiment analysis has a bright future ahead of it, especially on Twitter, and it's possible that a number of improvements will be made:

- Sentiment analysis will improve its comprehension of tweet context, including identifying sarcasm, irony, and subtle emotions.
- Improved language models will enable accurate analysis of tweets in several languages, expanding global sentiment insights.
- Sentiment analysis techniques provide real-time tracking, making it vital for businesses and organizations to respond quickly to public opinion.

Fig. 10: Heatmap

- Future sentiment analysis may connect with other AI domains, such as recommendation systems and predictive analytics, to provide more comprehensive insights for decision-making.
- Ethical Enhancements: As sentiment analysis advances, there will be a greater emphasis on privacy, openness, and eliminating biases in algorithms to ensure fair and ethical application.

4.7 Flowchart

Flowchart is shown in figure 11.

5 RESULT AND ANALYSIS

5.1 Data Overview

- **Tweet Collection:** A total of 75,682 tweets were collected on 14/08/2024. The tweets were gathered using keywords and hashtags related to gaming.
- **Preprocessing:** After data cleaning and preprocessing, 75,682 tweets were retained for analysis. The preprocessing steps included the removal of 0.01% of irrelevant content (e.g., URLs, special characters) and tokenization.
- **Count Vectorizer:** Count Vectorizer is a method that converts text into numeric values that a computer can comprehend. Here's how it works:
  - **Text Breakdown:** Every sentence or document is broken up into its component words.
  - **Word List Generation:** It compiles a list of all distinct words found in the text data.
  - **Word Frequency Counting:** It keeps track of how frequently each word from the list appears in a sentence or document.

  The number of features obtained is 311,037. We extracted the first 20 features.

5.2 Sentiment Distribution

Overall Sentiment (Through Pie Chart): The sentiment analysis classified the tweets into three categories: positive, negative, and neutral. The distribution of sentiments is as follows:

- **Positive Sentiment:** 38.7% (29,289 tweets)
- **Negative Sentiment:** 28.6% (21,645 tweets)
- **Neutral Sentiment:** 32.7% (24,748 tweets)

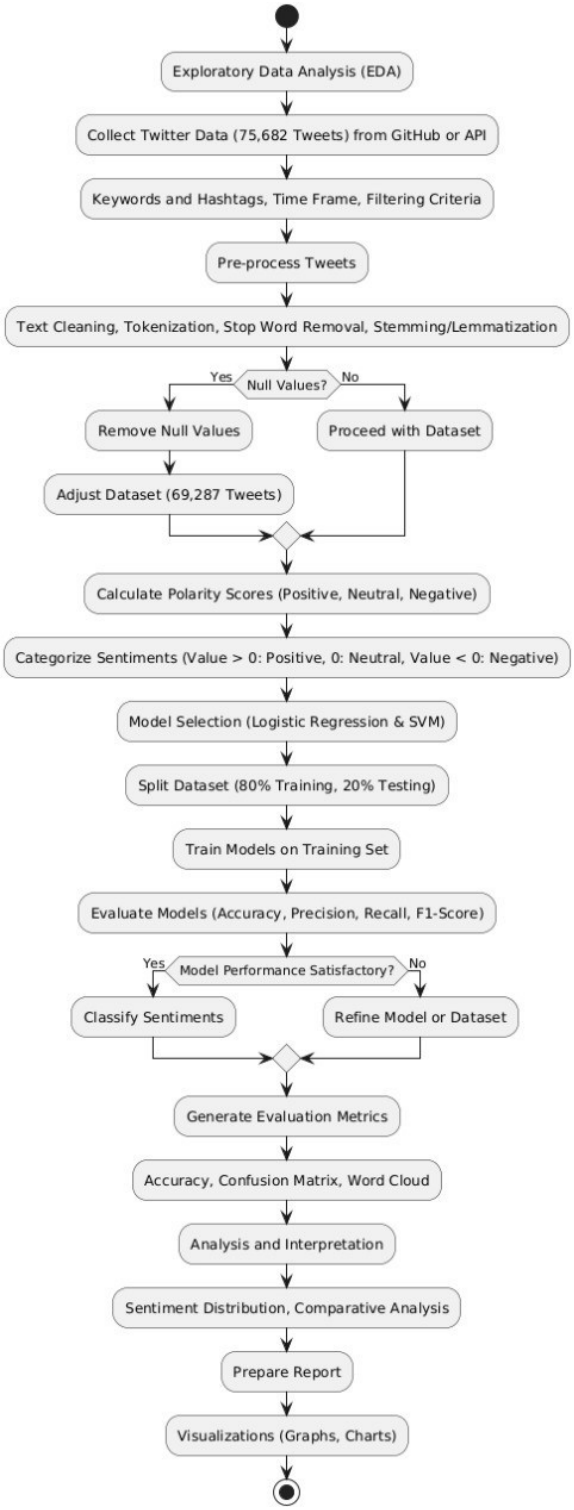

Fig. 11: Flowchart

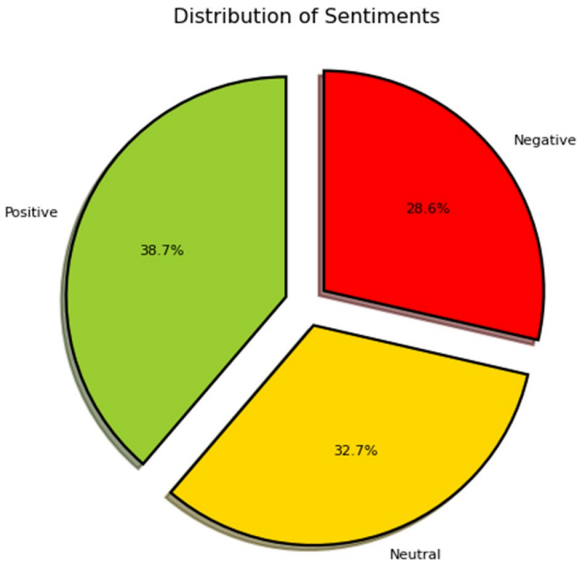

Fig. 12: Pie chart

Fig. 13: Logistic Regression

6 MODEL PERFORMANCE

6.1 Logistics Regression

- **Accuracy:** The sentiment analysis models achieved an overall accuracy of 96.75% in classifying the sentiments of the tweets.
- **The tuned Logistic Regression model:** achieved an accuracy of 96.95%, which is our overall tuned accuracy.
- **Precision, Recall, and F1-Score are shown in figure 13**

6.2 SVM

- **Accuracy:** The sentiment analysis models achieved an over- all accuracy of 97.49% in classifying the sentiments of the tweets.
- **The tuned Tuned SVM:** achieved an accuracy of 97.49%, which is our overall tuned accuracy.
- **Precision, Recall, and F1-Score are shown in figure 14**
- **Precision:**

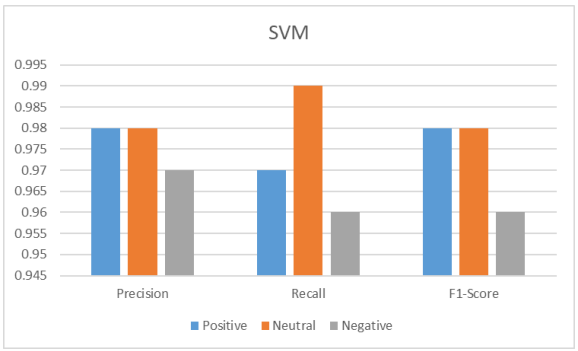

Fig. 14: SVM

- **Logistic Regression (First Chart):** Shows good precision in the *Positive* and *Neutral* categories, indicating that it correctly selects relevant examples for these sentiments.
- **SVM (Second Chart):** Also demonstrates high precision for *Positive* and *Neutral* categories, with slightly more stable alignment and values compared to Logistic Regression.

- **Recall:**
  - **Logistic Regression:** Recall is quite good for *Positive* and *Neutral*, but lower for *Negative*, suggesting that it may miss some important instances in the *Negative* class.
  - **SVM:** Recall is excellent for the *Neutral* category, outperforming logistic regression. However, it also underperforms in the *Negative* category, similarly to logistic regression.

- **F1-Score:**
  - **Logistic Regression:** F1-scores are balanced for *Positive* and *Neutral*, but the *Negative* class suffers due to lower recall.
  - **SVM:** Produces similar F1-score patterns, but with values that are more consistently aligned, especially in the *Neutral* category.

- **Conclusion:** Based on the results illustrated in the figure 14, SVM appears to have a slight advantage, particularly in recall and the overall alignment of the *Neutral* category. If your task prioritizes high recall (i.e., collecting as many relevant examples as possible) and consistency in *Positive* and *Neutral* classes, SVM is the better choice. However, if computational efficiency or model interpretability is more important, Logistic Regression remains a strong option. That said, based solely on these evaluation metrics, SVM slightly outperforms logistic regression in this classification task.

7 CALCULATION

7.1 Accuracy

The sentiment analysis in our SVM model achieved an overall accuracy of **97.48%** in classifying the sentiments of the tweets. This was the best model which obtained the highest accuracy.

$$\text{Accuracy} = \frac{\text{True Positive (TP)} + \text{True Negative (TN)}}{\text{Total Samples}} \tag{1}$$

$$\text{Accuracy} = \frac{3762 + 4455 + 5223}{13786} = \frac{13440}{13786} = 0.9749\,(97.49\%) \tag{2}$$

7.2 Precision

The percentage of accurately anticipated instances for each class is known as precision.

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}} \tag{3}$$

a) Negative Precision:

$$\text{Precision}_{\text{neg}} = \frac{3762}{3762 + 39 + 93} = \frac{3762}{3894} = 0.966\,(96.6\%) \tag{4}$$

b) Neutral Precision:

$$\text{Precision}_{\text{neu}} = \frac{4455}{49 + 4455 + 51} = \frac{4455}{4555} = 0.978\,(97.8\%) \tag{5}$$

c) Positive Precision:

$$\text{Precision}_{\text{pos}} = \frac{5223}{95 + 19 + 5223} = \frac{5223}{5337} = 0.979\,(97.9\%) \tag{6}$$

d) Weighted Precision:

$$\text{Weighted Precision} = \sum_i \left(\text{Precision}_i \times \frac{\text{Support}_i}{\text{Total Support}} \right) \tag{7}$$

$$\text{Weighted Precision} = \left(\frac{3773}{13786} \times 0.966 \right) + \left(\frac{4413}{13786} \times 0.978 \right) + \left(\frac{5254}{13786} \times 0.979 \right) \tag{8}$$

$$= 0.274 + 0.320 + 0.381 = 0.975\,(97.5\%) \tag{9}$$

7.3 Recall

The percentage of accurately predicted occurrences among actual cases for each class is known as recall.

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \tag{10}$$

a) Negative Recall:

$$\text{Recall}_{\text{neg}} = \frac{3762}{3906} = 0.963\,(96.3\%) \tag{11}$$

b) Neutral Recall:

$$\text{Recall}_{\text{neu}} = \frac{4455}{4513} = 0.987\,(98.7\%) \tag{12}$$

c) Positive Recall:

$$\text{Recall}_{\text{pos}} = \frac{5223}{5367} = 0.973\,(97.3\%) \tag{13}$$

d) Weighted Recall:

$$\text{Weighted Recall} = \sum_i \left(\text{Recall}_i \times \frac{\text{Support}_i}{\text{Total Support}} \right) \tag{14}$$

$$\text{Weighted Recall} = \left(\frac{3761}{13786} \times 0.478 \right) + \left(\frac{4454}{13786} \times 0.331 \right) + \left(\frac{5222}{13786} \times 0.091 \right) \tag{15}$$

$$= 0.273 + 0.323 + 0.379 = 0.975\,(97.5\%) \tag{16}$$

7.4 F1-Score

For every class, the F1-Score is the harmonic mean of Precision and Recall.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

a) Negative F1-Score:

$$\text{F1}_{\text{neg}} = \frac{2 \times 0.966 \times 0.963}{0.966 + 0.963} = \frac{1.860}{1.929} = 0.965\,(96.5\%) \tag{18}$$

b) Neutral F1-Score:

$$\text{F1}_{\text{neu}} = \frac{2 \times 0.978 \times 0.987}{0.978 + 0.987} = \frac{1.931}{1.965} = 0.982\,(98.2\%) \tag{19}$$

c) Positive F1-Score:

$$\text{F1}_{\text{pos}} = \frac{2 \times 0.979 \times 0.973}{0.979 + 0.973} = \frac{1.904}{1.952} = 0.976\,(97.6\%)$$

7.5 Hyperparameter Tuning in SVM

In hyperparameter tuning, we have used the following parameters: **C**, **Kernel**, **Degree**, and **Gamma**.

1. **C (parameter for regularization):**
   - **What it does:** C balances the decision boundary's complexity by managing the trade-off between achieving low training error and low testing error.
   - **High C value:** The model may overfit as it tries to classify every training point correctly.
   - **Low C value:** Allows some misclassification on the training set to form a simpler decision boundary and potentially improve generalization.
2. **Kernel:**
   - **What it does:** Specifies the type of decision boundary or how the SVM separates the data.
   - **Types:**
     - `"linear"`: Used for linearly separable data.
     - `"poly"`: The polynomial kernel; effective for non-linear data.
     - `"rbf"`: The Gaussian Radial Basis Function kernel; suitable for non-linear data.
     - `"sigmoid"`: A sigmoid kernel that mimics neural network behavior.
3. **Degree:**
   - **What it does:** Used only when `kernel='poly'`. It defines the degree of the polynomial function.

- **Higher degree:** Captures more complex patterns, but may lead to overfitting.
- **Lower degree:** Simpler decision boundary; less prone to overfitting but may underfit if the data is complex.

4. **Gamma:**
   - **What it does:** Determines how far the influence of a single training example reaches.
   - **High gamma:** Creates a more complex, wavy decision boundary; may lead to overfitting due to a smaller radius of influence per point.
   - **Low gamma:** Results in a smoother decision boundary; points have a wider radius of influence, which may lead to underfitting.

8 Conclusion

This study demonstrates how sentiment analysis can be a powerful technique for deriving insightful information from Twitter data to understand behavioral patterns, societal trends, and public opinion. To prepare the data for analysis, preprocessing techniques such as tokenization, stemming, and stop-word removal were applied to a dataset of 75,682 tweets. The distribution of the three sentiment categories—positive, neutral, and negative—was 38.7%, 32.7%, and 28.6%, respectively. To classify these sentiments, two machine learning models were trained: Support Vector Machines (SVM) and Logistic Regression. The SVM model outperformed Logistic Regression, achieving the highest accuracy of 97.48%, demonstrating its robustness for sentiment classification tasks.
Visualizations such as word clouds, pie charts, and confusion matrices provided valuable insights into sentiment trends and model performance. The study's findings highlight the potential of sentiment analysis across multiple fields. In marketing, businesses can use sentiment analytics to better understand consumer preferences and brand perceptions. In politics, analyzing public opinion helps strategists make informed decisions about policies and campaign tactics. In social behavior analysis, sentiment trends help measure societal opinions and concerns.
In particular, the use of SVM models illustrates the value of machine learning in addressing the challenges of unstructured social media data.
Despite its success, the study acknowledges several challenges, including difficulty in detecting sarcasm, handling multilingual datasets, and addressing biases in sentiment classification. Future work can overcome these limitations by incorporating context-aware language models such as transformers and leveraging advanced deep learning techniques. Furthermore, ensuring the ethical use of sentiment analysis—by safeguarding data privacy and minimizing algorithmic bias—remains essential for responsible implementation.
Sentiment analysis could yield even deeper insights when combined with other AI technologies like recommendation engines, predictive analytics, and real-time monitoring tools. This research lays the groundwork for more complex and scalable sentiment analysis systems, capable of delivering actionable insights across various domains and industries, especially in the context of growing demand for real-time public sentiment understanding.
Overall, this study demonstrates the effectiveness of various sentiment analysis methods in capturing public mood on Twitter and provides meaningful insights into the sentiment expressed in a large dataset of user comments. The results clarify whether comments on platforms like Twitter are professional or informal, and whether the sentiment conveyed is neutral, negative, or positive.

References

[1] Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media **8**, 216–225 (2014)

[2] Iglesias, C.A., Moreno, A.: Sentiment analysis for social media. Applied Sciences **9**(23), 5037 (2019)

[3] S. Jayasanka, E.M.I.A. T. Madhushani, Premaratne, S.: Sentiment analysis for social media. Unpublished (2013)

[4] L. Yue, X.L.e.a. W. Chen: A survey of sentiment analysis in social media. Knowledge and Information Systems **60**(3), 617–663 (2019)

[5] Prabowo, R., Thelwall, M.: Sentiment analysis: A combined approach. Journal of Informetrics **3**(2), 143–157 (2009)

[6] Kumar, A., Sebastian, T.: Sentiment analysis on twitter. International Journal of Computer Science Issues (IJCSI) **9**(4), 372 (2012)

[7] Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, pp. 627–666 (2010)

[8] Balahur, A.: Sentiment analysis in social media texts. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 120–128 (2013)

[9] Drus, Z., Khalid, H.: Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science **161**, 707–714 (2019)

[10] Grimalt-Alvaro, C., Usart, M.: Sentiment analysis for formative assessment in higher education: a systematic literature review. Journal of Computing in Higher Education **36**(3), 647–682 (2024)

[11] nez, M.R.-I., nez-Ventura, A.C., Castejón-Mateos, F., Cuenca-Jiménez, P.-M.: A review on sentiment analysis from social media platforms. Expert Systems with Applications **223**, 119862 (2023)

[12] Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: The Semantic Web – ISWC 2012, p. 32. Springer, ??? (2012)

[13] Sahayak, V., Shete, V., Pathan, A.: Sentiment analysis on twitter data. International Journal of Innovative Research in Advanced Engineering (IJIRAE) **2**(1), 178–183 (2015)

[14] Wankhade, M., Rao, A., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review **55**, 5731–5780 (2022)

[15] Singh, N., Tomar, D., Sangaiah, A.: Sentiment analysis: a review and comparative analysis over social media. Journal of Ambient Intelligence and Humanized Computing **11**(1), 97–117 (2020)

[16] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38 (2011)

[17] Mittal, A., Patidar, S.: Sentiment analysis on twitter data: A survey. In: Proceedings of the 7th International Conference on Computer and Communications Management, pp. 91–95 (2019)

[18] Huq, M., Ahmad, A., Rahman, A.: Sentiment analysis on twitter data using knn and svm. International Journal of Advanced Computer Science and Applications **8**(6) (2017)

[19] Desai, M., Mehta, M.: Techniques for sentiment analysis of twitter data: A comprehensive survey. In: 2016 International Conference on Computing, Communication and Automation (ICCCA), pp. 149–154 (2016)

[20] Wongkar, M., Angdresey, A.: Sentiment analysis using naive bayes algorithm of the data crawler: Twitter. In: Proceedings of the International Conference on Information and Communication Technology, pp. 1–5 (2019)

[21] Dattu, B., Gore, D.: A survey on sentiment analysis on twitter data using different techniques. International Journal of Computer Science and Information Technologies **6**(6), 5358–5362 (2015)

[22] Adwan, O., Al-Tawil, M., Huneiti, A., Shahin, R., Zayed, A.A., Al-Dibsi, R.: Twitter sentiment analysis approaches: A survey. International Journal of Emerging Technologies in Learning (iJET) **15**(15), 79–93 (2020). Retrieved October 26, 2024 from https://www.learntechlib.org/p/217980/

[23] Alsaeedi, A., Khan, M.: A study on sentiment analysis techniques of twitter data. International Journal of Advanced Computer Science and Applications **10**(2), 361–374 (2019)

[24] Mehta, P., Pandya, S.: A review on sentiment analysis methodologies, practices and applications. International Journal of Scientific and Technology Research **9**(2), 601–609 (2020)

[25] Singh, S., Paul, S., Kumar, D., Arfi, H.: Sentiment analysis of twitter data set: survey. International Journal of Applied Engineering Research **9**(22), 13925–13936 (2014)