Predicting User Sentiments in Social Media with Machine Learning and Natural Language Processing techniques

Kurmala Lakshmi Sathvika[1], Devineni Srujitha[1], Kavuru Tanya[1], Srilatha Tokala[1], Murali Krishna Enduri[1], Satish Anamalamudi[1]

[1] Algorithms and Complexity Theory Lab, Department of Computer Science and Engineering, SRM University-AP, Neerukonda, Andhra Pradesh, India.

Contributing authors: sathvika_kurmala@srmap.edu.in ; srujitha_devineni@srmap.edu.in ; tanya_kavuru@srmap.edu.in ; srilatha_tokala@srmap.edu.in; muralikrishna.e@srmap.edu.in; satish.a@srmap.edu.in;

Abstract

Social media platforms have become an integral part of modern communication, offering individuals an unprecedented opportunity to express their thoughts, opinions, and emotions publicly. The primary objective of social media trends sentiment analysis is to analyze and interpret the sentiments associated with specific hashtags, topics, or viral content through the application of natural language processing and machine learning algorithms. To analyze the trends, we have used various ML algorithms such as CNN, Naive Bayes, SVM, Random-forest classifier, and Linear Regression. We have analyzed these algorithms by calculating various factors like f1 score, recall, precision, and accuracy. As social media platforms continue to evolve, this research field presents exciting opportunities for businesses, researchers, and policymakers to harness the collective voice of the online community and stay informed about the ever-changing public sentiment in the era of digital communication.

Keywords: Machine learning algorithms; Natural Language Processing; Text Mining; Twitter; Lexicon Based Methods

1 Introduction

Social media is one of the many platforms that involves and considers the emotions and opinions of people all over the world. Social media sentiment analysis plays a pivotal role in extracting sentiments or opinions out of the content posted on various platforms like Twitter, Facebook etc [1]. The emergence of social media, access to opinions has become considerably more manageable and it's more critical than ever to measure social sentiment, as it changes frequently. Through this we can judge whether the posts are positive, negative, or neutral. We can analyze these by collecting the datasets and using the Machine Learning algorithms accordingly to classify various parameters [2]. Using this technique, businesses, governments, and individuals tend to understand public opinion, identify ongoing trends, and make informed decisions. The study in this field is crucial due to several compelling reasons like it can be used in evaluating your respective brand's health, dealing with a crisis, understanding the competition, social listening and trend analysis, and customer feedback. Overall, sentiment analysis plays a pivotal role in extracting valuable insights from vast amounts of textual data generated on social media platforms or any other digital platforms [3]. It empowers researchers, businesses, and policymakers to make informed decisions, enhance user experiences, and engage with the public more efficiently.

ML and Deep Learning techniques offer valuable tools for exploring and understanding sentiment analysis. By leveraging diverse data sets, extracting meaningful features from the data, and utilizing various learning algorithms, these techniques can aid trend analysis, public opinion, customer service, and support [4]. Social media sentiment analysis is a specialized application of natural language processing (NLP) and machine learning techniques to analyze and interpret sentiments, emotions, and opinions. With the explosive growth of social media, this has become a crucial tool for understanding public perception, gauging customer satisfaction, and making data-driven decisions in diverse fields. Researchers have been using various methods such as lexicon-based approach, and hybrid approach, which is a way to get observations and draw conclusions [5]. Now in the modern era after the advent of ML and DL algorithms, numerous ways have evolved to efficiently collect, store, and analyze the data and perform complex computations. We are trying to solve this problem by identifying the patterns within the data extracting the pattern and calculating the trends. The study is important because it serves as a valuable resource for those seeking to leverage sentiment analysis to gain insights into public sentiment and opinion in the era of social media. This study also delves into preprocessing techniques used to handle challenges specific to social media data. The overarching aim is to gain valuable insights into public perception, attitudes, and reactions towards various topics, brands, events, and trends in the digital realm. In sentiment analysis of social media, researchers and practitioners observe several knowledge patterns that emerge from the analysis of textual data and the application of machine learning techniques [6]. These knowledge patterns help in understanding the sentiments, emotions, and opinions expressed by users on social media platforms. Some of the common knowledge patterns observed in sentiment analysis of social media are Viral sentiment Cascades, Sentiment Polarity Distribution, Emotional Analysis, and Contextual Sentiment.

2 Related Works

Becker *et al.* proposed an innovative event identification method in social media using ensemble learning [7]. They assess text, time, and location, and introduce "URLs" and "bursty vocabularies" as new factors. Their evaluation of the upcoming dataset focused on Flickr photos preceding events demonstrates high accuracy. Challenges include prolonged inactivity in large-scale events and evolving bursty vocabulary, suggesting potential improvements with user participation and word frequency considerations. Gautam *et al.*, analysis for classifying customer reviews, which is useful for analyzing data in the form of the number of tweets with very illogical and either favorable or negative thoughts has been discussed [8]. The user's opinion has been classified into various sentiment classes which further help in the decision-making process. Sentamiselvan *et al.*, they analyzed the movie reviews given by people on various social media platforms through tweets, blogs, etc. using various techniques [9]. With this, they have identified the sentiment of an individual with respect to a given source of information. In the study by Ain *et al.*, enormous heaps of data from social networks, forums, and review sites have been gathered and analyzed to calculate the sentiment analysis [10]. To solve the problem of insufficient labeled data in the field of NLP, deep learning techniques are merged due to their automatic learning capability. It highlights studies that involve the implementation of convolutional neural networks, and deep neural networks. Jagdale *et al.*, A computer analysis of a person's buying interests and opinions about a company's business entity is also done in order to better business strategy and obtain insight into customer comments about their items [11].

In their investigation, Yue *et al.* focuses on providing standard approaches in the field of sentiment analysis from three distinct viewpoints [12]. Particularly, several different strategies and methodologies are compared and categorized. On the other hand, several data formats and cutting-edge research methods are introduced, along with their limits. These resources serve as the foundation for identifying and discussing the key future prospects for sentiment analysis.

The research by Nilekar *et al.* collected Twitter data, focused on text, and used TF-IDF and machine learning to achieve 85% accuracy in trending tweet extraction, 80% in topic classification, and 75% in sentiment analysis for Arabic tweets [13]. Their approach advances trend analysis, offering the potential for business and organizational use in tracking opinions and trends. In this, Agboola *et al.* address challenges in spam filtering with a machine learning approach [14]. They use word embedding to convert words into vectors and apply various algorithms, achieving an impressive 96% accuracy on a spam message dataset. This underscores machine learning's effectiveness in adaptable spam detection, overcoming evolving spammer tactics. The next section describes about different machine learning methods that are used in our analysis.

3 Methodology

In this section, we outline the methods employed for conducting sentiment analysis on Twitter datasets. The methodology encompasses data preprocessing, feature extraction, model building, and evaluation using a range of machine learning algorithms.

The following steps were taken to achieve accurate sentiment classification as shown in Figure 1.

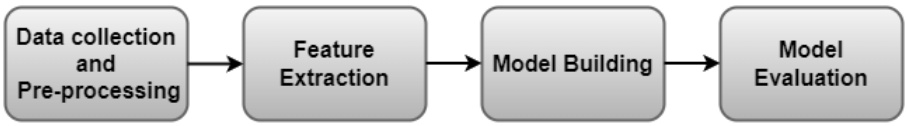

Fig. 1 Sentiment Analysis Process.

Data Collection and Preprocessing: We obtained Twitter datasets from Kaggle, that provided the necessary information for conducting sentiment analysis. We followed the subsequent steps for data preprocessing:
Data Cleaning: The datasets were cleaned to remove irrelevant information and focus on the essential attributes. Null values were either dropped or replaced based on the specific dataset and attribute.
Tokenization and Stemming: We employed the Porter Stemmer algorithm for stemming, which involved converting words into their root forms. Tokenization was performed to split tweets into individual words for analysis.
Removal of User Handles: Twitter user handles (starting with "@") were removed, as they were not informative for sentiment analysis.
Feature Extraction: We extracted features from the cleaned and processed tweets using various techniques. The following methods were employed:
Bag-of-Words: We applied the Bag-of-Words technique to convert the tokenized tweets into numerical vectors. Each word in the vocabulary was treated as a feature, and its frequency in each tweet was recorded [15].
TF-IDF (Term Frequency-Inverse Document Frequency): We used TF-IDF to calculate the importance of words in each tweet relative to the entire dataset. This technique considered both term frequency and the inverse document frequency to capture the significance of words.
Model Building: We employed a range of machine learning algorithms to build sentiment analysis models on the training dataset. The models were then evaluated on the testing dataset to assess their performance. The algorithms used were:
SVM algorithm: In sentiment analysis, each text document is represented as a high-dimensional feature vector, where each feature represents a word or n-gram. SVMs can handle this high-dimensional feature space efficiently. They are used for sentiment analysis because they are effective in handling high-dimensional text data, can capture non-linear relationships between features, are robust to overfitting, can handle imbalanced datasets, and have a solid theoretical foundation [16].
KNN algorithm: It is not a common choice for sentiment analysis as it is mostly for regression tasks [17]. But is primarily a classification algorithm which makes it a good choice and also a complex choice for sentiment analysis. This is employed for sentiment analysis in the case of simple scenarios to understand its behavior and limitations when compared to other algorithms. If we ever need a quick and simple sentiment analysis for a small dataset, this algorithm could be implemented without extensive

model building or for feature engineering.
Logistic-Regression: Logistic Regression is widely used, and also a common choice for sentiment analysis. The goal is to determine the sentiment expressed in a piece of text. This algorithm provides results that are interpretable. The coefficients of this model could be analyzed to understand which words and features contribute positive, negative, and neutral emotions to the sentiment prediction. This algorithm is also computationally efficient and it requires less computational resources, unlike more complex models. These algorithm coefficients help us to identify important features that contribute to a particular sentiment. This even works well with relatively small datasets, which are common in sentiment analysis tasks. *Decision Tree Classifier:* This algorithm mainly consists of a Tree structure with nodes, which is an important part of explaining this algorithm. Those are called Tree nodes and root nodes, where each node represents the decision or a test on an input feature. These nodes are connected by branches. Root nodes are at the top of the tree structure and Leaf nodes are the final nodes of the tree that represent output or prediction of a particular subset of data. This algorithm is known for its interpretable nature and can handle both categorical and numerical types of data. *Naive Bayes:* This is a probabilistic classifier that is most used in Natural Language Processing techniques like sentiment analysis. In Naive Bayes, probabilities are assigned to words or phrases, segregating them into different labels. Comparatively, this algorithm is much faster as it calculates the probabilities. Its easily scalable property makes it the most efficient algorithm.
CNN(Convolutional Neutral Networks): CNNs excel at capturing local patterns or features within the input data. In the context of sentiment analysis, local patterns can represent important clues about the sentiment expressed in a text. CNNs can learn hierarchical representations of the input data.
XG Boost: Basically, it's not a choice for sentiment analysis, as it's usually used for structured data problems [18]. But it's still possible to make XG Boost adapt for sentiment analysis tasks, and it will do its work efficiently. If the sentiment analysis task you provided involves both text data and structured features, XG Boost can handle the structured part efficiently while incorporating text features. XG Boost is widely known for use in hybrid models where you combine predictions from different models. For instance, the NLP model can be useful for text analysis and XG Boost for structured features as discussed, and then getting accuracy by combining their predictions.
Model Evaluation: The performance of each model was evaluated using key metrics, including accuracy, precision, recall, and F1-score.

The methodology outlined above facilitated the successful execution of sentiment analysis on Twitter datasets. By employing data preprocessing, feature extraction, and a diverse set of machine learning algorithms, we achieved accurate sentiment classification for the tweets. In this section, we will further discuss the characteristics of the datasets used in our study, shedding light on their relevance and significance. and the subsequent analysis of the sentiment patterns found in the Twitter data.

4 Dataset Statistics

In this study, the foundation of our research lies in three distinctive datasets meticulously gathered from Kaggle: "EDA For Sentiment Analysis," "EDA Twitter Sentiments," and "Twitter Sentiment Dataset." These datasets form the bedrock of our exploration into sentiment analysis, providing the essential raw materials for our analytical journey. Let's delve into the specifics of each dataset:

Table 1 Dataset Statistics

Dataset	**No of rows**	**No of columns**	**Attributes**	**Text Sentiment**
EDA Twitter Sentiments	27481	3	Id, Text, Sentiment	Positive, Negative
Twitter Sentiments	31962	3	Id, Label, Tweet	Positive, Neutral, Negative
EDA for Sentiment Analysis	10000	2	Clean_text, category	Positive, Neutral, Negative

These datasets serve as the pillars of our study, providing a diverse range of textual expressions harnessed from social media platforms. Our analysis hinges on the comprehensive evaluation of these datasets, each contributing a unique dimension to our understanding of sentiment within various contextual frameworks. By dissecting the attributes and sentiments embedded within these datasets, we aim to construct robust sentiment analysis models capable of transcending the confines of any single dataset. This approach ensures that our research is adaptable and versatile, effectively addressing real-world sentiment dynamics across diverse scenarios.

5 Results

Having discussed the significance of the datasets, we will now delve into the outcomes of our sentiment analysis study. The performance of numerous machine learning algorithms across distinct datasets is examined in this study, illuminating their effectiveness in sentiment categorization. The evaluation encompassed accuracy, precision, recall and f1-score integral metrics that collectively illuminate the strengths and adaptability of these algorithms.

Let us consider M as True Positives - Instances correctly predicted as positive sentiments. N as False Positives - Instances incorrectly predicted as positive sentiments. O as False Negatives - Instances incorrectly predicted as negative sentiments when they were actually positive. P as True Negatives - Instances correctly predicted as negative sentiments.

Accuracy: Accuracy gauges the proportion of correct predictions in the entire dataset, reflecting how effectively models predict sentiment polarity from text data. It encompasses both true positive and true negative predictions and is calculated using the following formula:

$$Accuracy = \frac{m + p}{(m + n + o + p)}$$

Precision: Precision provides insights into the reliability of positive and negative sentiment predictions. It measures the proportion of true positive predictions (m) to all instances predicted as positive (m + n):

$$Precision = \frac{m}{(m+n)}$$

Recall: Recall measures a model's ability to correctly identify all instances of a particular sentiment class. It's the ratio of true positive predictions (m) to the total instances of that class (m + o):

$$Recall = \frac{m}{(m+o)}$$

F1 Score: The F1 score balances precision and recall, especially valuable for imbalanced datasets. It's defined as:

$$F1Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

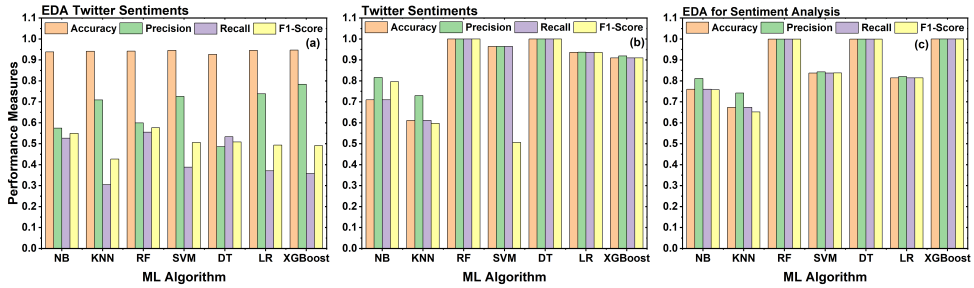

Fig. 2 Performance analysis on three datasets namely EDA Twitter Sentiments, Twitter Sentiment, and EDA For Sentiment Analysis.

Figure 2 (a) steers our focus to the EDA Twitter Sentiments dataset, which encompasses Neutral sentiments, we extend our analytical purview. Introduces a graphical dimension, presenting a visual depiction of algorithmic performances. Here, the XGBoost algorithm maintains its dominance, securing the highest accuracy of 0.947. Notably, the Logistic Regression and SVM method match closely with an almost impeccable accuracy of 0.945, reaffirming its robust standing. Nevertheless, in contrast, the Decision Tree algorithm surfaces with the lowest accuracy of 0.926. Expanding our analysis in Figure 2 (b) is the Twitter Sentiment dataset, aligning with characteristics of the prior dataset, we encounter a distinct algorithmic panorama. A visual representation further enhances our grasp of these algorithm's performances. Here, the Random forest and Decision tree algorithm emerges as a beacon of accuracy, boasting a pristine 1.0. Concurrently, the XGBoost method maintains exceptional accuracy at 0.909. Precision, recall, and f1-score harmoniously orbit accuracy. Nevertheless, in

tandem with previous datasets, the KNN algorithm persistently reports the lowest accuracy of 0.610. Figure 2 (c) deals with the analysis of the EDA For Sentiment Analysis dataset ventured into a rich tapestry of XGBoost algorithm with higher accuracy, precision, recall, and f1-score of 1.0. As depicted in the figure, a visual representation adds depth to our understanding of each algorithm's performance on this dataset. Foremost among them, the Decision Tree and Random forest distinguished itself with the second highest accuracies with 0.99. In contrast, the KNN algorithm exhibited the dataset's lowest accuracy of 0.673.

In assessing algorithm performance across three sentiment analysis datasets, the XGBoost algorithm consistently emerged as the top performer, showcasing the highest accuracy, precision, recall, and f1-score in all cases. The decision tree and random forest algorithm demonstrated its strength as the second-best performer, with consistently high accuracy and commendable precision, recall, and f1- scores. The SVM algorithm also proved competitive, securing a solid third position. However, the k-nearest neighbors algorithm consistently reported the lowest accuracy, highlighting its limitations in sentiment analysis. Overall, these results emphasize the dominance of XGBoost, decision tree, random forest, and SVM while reinforcing the significance of algorithm selection in achieving accurate sentiment classification. Our study stands as a repository of enlightening algorithmic insights, revealing the intricate interplay between algorithms and datasets. Empowered with this comprehension, decision-makers are better equipped to orchestrate algorithmic selections that yield optimal results in diverse scenarios.

6 Conclusion and Future Work

In this sentiment analysis of the twitter datasets, we aimed to analyze and predict the overall sentiment expressed by users on the social media platform. We have utilized NLP techniques to interpret and summon a large number of tweets. We observed and calculated positive, negative, and neutral sentiments for the twitter datasets. As twitter consists of hashtags and mentions, we also have inspected and performed exploratory data analysis. With the help of ML algorithms, we have calculated the precision, accuracy, f1-score, and recall. With this, we were able to try and test various algorithms and conclude the best algorithms taking into count various parameters. Overall, the sentiment analysis of twitter or any social media helps us to provide insights into public opinion, and emotional trends which makes it a powerful tool for understanding and reviewing the sentiment of a diverse community on social media. When used responsibly and in conjunction with other research methods, sentiment analysis can offer meaningful and actionable insights for various fields, including market research, social studies, and public opinion analysis.

Currently, the study can lead to gaining the attention of the respective authorities that can address the issues which are obtained during the analysis, and in the future the obtained results can be used to deploy a perfect model. We are interested to explore and examine other datasets and try new social media platforms rather than Twitter. Future work for this sentiment analysis can also focus on advancing the existing techniques and addressing the challenges and limitations observed in current

approaches. Some potential areas of improvement and expansion may include fine-grained sentiment analysis in which we will enhance the sentiment analysis models by distinguishing various emotions such as joy, fear, sadness, etc., and multilingual sentiment analysis- this includes extending sentiment analysis models to support multiple languages. Twitter is a global platform with diverse users, and sentiment analysis should be adaptable to different languages, and sentiment over time which includes the evolution of sentiment trends over longer periods to identify patterns and understand how sentiment changes in response to global events and societal shifts.

References

[1] Pang, B., Lee, L., *et al.*: Opinion mining and sentiment analysis. Foundations and Trends in information retrieval **2**(1–2), 1–135 (2008)

[2] Patel, V., Prabhu, G., Bhowmick, K.: A survey of opinion mining and sentiment analysis. International Journal of Computer Applications **131**(1), 24–27 (2015)

[3] Philander, K., Zhong, Y.: Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. International Journal of Hospitality Management **55**, 16–24 (2016)

[4] Rout, J.K., Choo, K.-K.R., Dash, A.K., Bakshi, S., Jena, S.K., Williams, K.L.: A model for sentiment and emotion analysis of unstructured social media text. Electronic Commerce Research **18**, 181–199 (2018)

[5] Choudrie, J., Patil, S., Kotecha, K., Matta, N., Pappas, I.: Applying and understanding an advanced, novel deep learning approach: A covid 19, text based, emotions analysis study. Information Systems Frontiers **23**, 1431–1465 (2021)

[6] Ibrahim, N.F., Wang, X.: A text analytics approach for online retailing service improvement: Evidence from twitter. Decision Support Systems **121**, 37–50 (2019)

[7] Becker, H., Naaman, M., Gravano, L.: Event identification in social media. In: WebDB (2009)

[8] Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In: 2014 Seventh International Conference on Contemporary Computing (IC3), pp. 437–442 (2014). IEEE

[9] Sentamilselvan, K., Aneri, D., Athithiya, A., Kumar, P.K.: Twitter sentiment analysis using machine learning techniques. International Journal of Engineering and Advanced Technology (IJEAT) **9**(3), 1–9 (2020)

[10] Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., Rehman, A.: Sentiment analysis using deep learning techniques: a review. International Journal of Advanced Computer Science and Applications **8**(6) (2017)

[11] Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N.: Sentiment analysis on product reviews using machine learning techniques. In: Cognitive Informatics and Soft Computing: Proceeding of CISC 2017, pp. 639–647 (2019). Springer

[12] Yue, L., Chen, W., Li, X., Zuo, W., Yin, M.: A survey of sentiment analysis in social media. Knowledge and Information Systems **60**, 617–663 (2019)

[13] Nilekar, S., Rawat, S., Verma, R., Rahate, P.: Twitter trend analysis. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol. (2020)

[14] Agboola, O.: Spam detection using machine learning and deep learning. PhD thesis, Louisiana State University and Agricultural & Mechanical College (2022)

[15] Neogi, A.S., Garg, K.A., Mishra, R.K., Dwivedi, Y.K.: Sentiment analysis and classification of indian farmers' protest using twitter data. International Journal of Information Management Data Insights **1**(2), 100019 (2021)

[16] Krawczyk, B., McInnes, B.T., Cano, A.: Sentiment classification from multi-class imbalanced twitter data using binarization. In: Hybrid Artificial Intelligent Systems: 12th International Conference, HAIS 2017, La Rioja, Spain, June 21-23, 2017, Proceedings 12, pp. 26–37 (2017). Springer

[17] Imandoust, S.B., Bolandraftar, M., *et al.*: Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International journal of engineering research and applications **3**(5), 605–610 (2013)

[18] Si, M., Du, K.: Development of a predictive emissions model using a gradient boosting machine learning method. Environmental Technology & Innovation **20**, 101028 (2020)