International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Research Article)

Check for updates

Machine learning-based text analysis system for multilingual data

Sameeksha Yadav [1], Garima Srivastava [1,*] and Lalita Kumari [2]

[1] *Department of Computer Science and Engineering, Amity University Uttar Pradesh, India.*
[2] *Department of Computer Science and Engineering, Amity University Patna, India.*

International Journal of Science and Research Archive, 2026, 19(02), 773-778

Publication history: Received on 27 March 2026; revised on 09 May 2026; accepted on 11 May 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.2.1052

Abstract

In the era of social media, Twitter stands out as a rich source of opinions. Choosing the topic of Twitter sentiment analyzer was driven by the pressing need to comprehend the ever-evolving sentiment dynamics. The significance lies in harnessing the power of social media data to gain insights into public opinion and emerging trends, thereby aiding businesses and policymakers in making informed decisions. Despite the strides made in sentiment analysis, there remain gaps in existing research. Many models struggle with nuanced expressions, context-dependent sentiments, and the evolving language on social media platforms. Addition- ally, the challenges of handling sarcasm and detecting sentiment shifts within a single tweet pose significant hurdles. Addressing these gaps, our approach involves the integration of advanced natural language processing techniques and context-aware sentiment analysis methods. By delving into the intricacies of linguistic nuances and context, we aim to enhance the accuracy of sentiment predictions on Twitter.

Keywords: Sentiment Analysis; Twitter Data; Emotional Intelligence; Data labeling

1. Introduction

Twitter has become an influential channel for social media analytics. Its sentiment analysis is crucial for grasping the intricacies of online interactions. The sentiment analysis technique on Twitter is an extensive framework that surpasses basic tweet categorizations as positive, negative, or neutral. It delves deeper into nuanced emotions like joy, anger, sadness, or surprise, along with the intensity and context behind them.

It utilizes sophisticated Natural Language Processing (NLP) techniques to discern subtleties such as sarcasm, irony, or slang, thereby enhancing the precision and depth of sentiment classification. One of the significant advantages of Twitter sentiment analysis is that it facilitates the tracking of sentiment trends over time, enabling the identification of emerging patterns, shifts, or spikes linked to specific events or announcements.

Sentiment analysis on Twitter isn't just about words—it also includes pictures and videos. By looking at both words and visuals, we get a better under- standing of how people express their emotions on Twitter. This helps us learn more about sentiment on the platform and opens up new possibilities for studying things like storytelling in pictures, how brands are seen, and keeping content appropriate.

Various studies have been conducted on sentiment analysis on Twitter to determine the emotions expressed in tweets. Different approaches have been tried, such as analyzing the words used, their structure, and the timing of the tweet, to determine if the sentiment is positive, negative, or neutral. Researchers have also examined how factors like age, language, and timing can influence how people express their feelings on social media.

* Corresponding author: Garima Srivastava

Sentiment analysis on Twitter has been found to have several benefits, including the ability to monitor people's attitudes toward brands, comprehend current trends, and observe people's reactions during emergencies. By building upon prior research, new studies can further enhance sentiment analysis and explore novel ways of using it to gain a better understanding of online occurrences.

The purpose of this paper is to elucidate the importance of sentiment analysis on Twitter and how it can be leveraged to comprehend people's sentiments and emotions on the internet. Twitter is a vast platform where people express their opinions and reactions on various topics. By analyzing tweets, we can determine whether people have a positive, negative, or neutral viewpoint about something. This information is beneficial for businesses, governments, and researchers to make informed decisions and comprehend what is occurring worldwide.

The paper will discuss various methods to analyze tweets and demonstrate how this data can be applied in real life, such as for marketing or understanding public opinion. Additionally, it aims to propose innovative ideas to enhance sentiment analysis on Twitter. Overall, this paper intends to emphasize why sentiment analysis on Twitter is valuable and how it can benefit everyone.

2. Literature Review

Several salient characteristics highlighted in the surveyed reports are:

The author's review addresses the challenges and motivations in Twitter Sentiment Analysis (TSA)[1], focusing on the brevity of tweets, sentiment class imbalance, and stream-based tweet generation. It discusses a benchmark evaluation of 28 TSA systems, revealing an average accuracy of 61% and highlighting the superiority of domain-specific approaches. The review emphasizes error analysis, unveiling 13 misclassification categories, and explores the interplay between sentiment classification performance and event detection effectiveness.

The author investigates the impact of diverse text preprocessing methods on sentiment analysis across multiple Twitter datasets[2]. Key findings include the minimal effect of URL removal, varying sensitivities of classifiers to stop words and numbers, positive outcomes from expanding acronyms, and the significance of replacing negations.

Using crowdsourcing and thorough preprocessing, the study presents an effective real-time sentiment assessment method for the 2012 U.S[3]. Presidential Election on Twitter, achieving a 59% accuracy rate.

The author suggested that GloVe-DCNN performs exceptionally well in Twitter sentiment analysis, outperforming baseline techniques with its ability to capture contextual details[4].

The author suggested that adding semantic elements improves accuracy, although limitations remain in entity specificity and contextual understanding[5].

Machine learning algorithms such as SVM, Maximum Entropy, and Naıve Bayes have been used effectively for sentiment analysis on Twitter[6].

Studies show that analyzing healthcare-related tweets helps understand public opinion but requires domain-specific tools[7]. Machine learning-based sentiment analysis systems provide automation in analyzing large volumes of social media data[8].

Twitter sentiment analysis plays a role in policy-making and understanding public opinion during major events[9]. Godbole et al. (2007) discussed large-scale sentiment analysis for news and blogs[10]. Davidov et al. (2010) improved sentiment learning using hashtags and emoticons[11]. Ghiassi et al. (2013) proposed hybrid systems combining neural networks and n-gram models[12].

Hu and Liu (2004) worked on customer review sentiment analysis[13]. Barbosa and Feng (2010) addressed bias and noise in Twitter sentiment analysis[14]. Thelwall et al. (2012) enhanced sentiment analysis techniques for social media[15]. Bollen et al. (2011) studied the relationship between Twitter sentiment and stock market trends[16].

2.1. Dataset

The dataset that we have taken is Sentiment140. In this dataset, 1,600,000 tweets were retrieved using the Twitter API. Sentiment analysis has been performed on these tweets wherein each tweet is annotated with 0 as negative and 1 as positive.

In the initial stage, the Kaggle library was installed and configured to access the dataset. After that, the dataset was downloaded using the API command and the compressed file was extracted successfully.

Further, the required libraries such as NumPy and Pandas were imported. NumPy was used for numerical computation, while Pandas was used for data manipulation and analysis.

Stopwords were downloaded using the Natural Language Toolkit (NLTK), which helps in removing commonly used words such as ”the”, ”is”, and ”and” that do not contribute significantly to sentiment analysis.

The dataset was then loaded into a Pandas dataframe by reading the CSV file. Column names were assigned, and the initial rows of the dataset were printed to verify that the data was loaded correctly.

2.2. Data Preprocessing

In our research, data preprocessing is an essential stage that is designed to maximize the sentiment analysis potential of the Twitter dataset. Utilizing the Twitter API connection, we carefully gathered tweets, which allowed us to create a real-time and diverse dataset that captures user sentiments effectively.

Initially, the dataset was examined to identify missing values. After a thor ough analysis, it was observed that there were no missing values present in the dataset, ensuring data completeness and reliability.

Next, the distribution of the target column was analyzed to understand the proportion of positive and negative sentiments. This step is important to evaluate whether the dataset is balanced or not.

Further, text preprocessing techniques were applied. Stemming was per- formed to reduce words to their root form, which helps in minimizing redundancy and improving model efficiency. After stemming, a new column containing the processed text was added to the dataset.

The dataset was then divided into two parts: training data and testing data. The training data is used to train the machine learning model, while the testing data is used to evaluate its performance.

Finally, vectorization was applied to convert textual data into numerical for- mat. This step is essential because machine learning algorithms cannot process raw text directly. Vectorization transforms text into a structured numerical rep-resentation, enabling efficient computation and analysis.

3. Methodology

To perform sentiment analysis on Twitter data using Python, we followed a structured approach consisting of several steps.

Firstly, the dataset was accessed from the Kaggle library using the API command. This allowed us to obtain a large-scale dataset efficiently for analysis.

Next, essential libraries such as NumPy and Pandas were imported for data manipulation and analysis. The dataset was then cleaned by removing unnecessary elements such as stopwords and irrelevant text.

After cleaning, stemming was applied to reduce words to their root forms, which helps in improving the consistency of the data. The dataset was also checked for missing values to ensure data quality.

The processed dataset was then divided into two parts: training data and testing data. The training dataset was used to build the machine learning model, while the testing dataset was used to evaluate its performance.

Further, vectorization techniques were applied to convert textual data into numerical form. This step is necessary because machine learning models require numerical input for processing.

Finally, a logistic regression model was applied for sentiment classification. Logistic regression is a statistical method used for binary classification tasks. It predicts the probability of a tweet belonging to a particular sentiment class (positive or negative) using a logistic (sigmoid) function. This model helped in analyzing the sentiment expressed in the tweets and provided an effective prediction of sentiment polarity.

4. Results

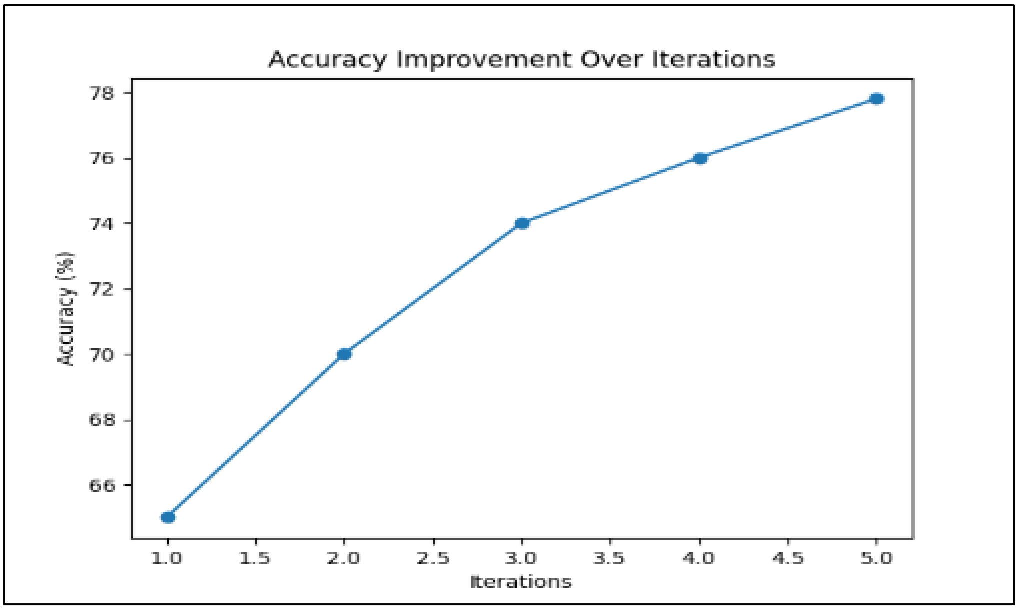

Figure 1 Model accuracy over iterations

This study analyzes the effectiveness of machine learning in performing sentiment analysis on Twitter data. The dataset was collected using the Twitter API, ensuring that it was both extensive and relevant to the objectives of the research. Before applying the model, the data was carefully preprocessed[18]. This included removing unnecessary elements such as URLs and special characters, performing tokenization to break text into individual words, and eliminating stopwords that do not contribute significantly to sentiment analysis. Additionally, normalization was performed by converting all text into lowercase to maintain consistency.

The dataset was divided into two subsets: a training set and a testing set. The training set was used to train the model, while the testing set was used to evaluate its performance. This approach helps in avoiding overfitting, where a model performs well on known data but poorly on unseen data.

A logistic regression model was applied to classify the sentiment of tweets. To evaluate the performance of the model, several metrics were considered, including accuracy, precision, recall, and F1-score[19][20][21].

The results show that the logistic regression model achieved an accuracy of 77.8%. This demonstrates that machine learning techniques can effectively analyze sentiment in Twitter data and provide meaningful insights from large- scale social media content.

5. Conclusion

The development of a sentiment analysis system tailored for the dynamic Twitter environment was the focus of this study. The system effectively classifies the sentiment represented in tweets as positive or negative by utilizing machine learning techniques. This provides valuable insights into public opinion on a large scale.

The study highlights the significance of sentiment analysis in real-world applications. Businesses can use this information to improve their strategies and better understand customer preferences. Similarly, policymakers can analyze public opinion on important issues to make informed decisions.

However, the study also acknowledges certain challenges in Twitter sentiment analysis. These include the short length of tweets, imbalance in sentiment distribution, and the difficulty of interpreting sarcasm and complex language.

Despite these challenges, the techniques used in this research, including data preprocessing and feature engineering, provide a strong foundation for sentiment analysis. Future work can focus on improving model accuracy by using advanced machine learning and deep learning techniques, along with better context understanding and sarcasm detection.

Overall, this study demonstrates the potential of sentiment analysis as a powerful tool for understanding public opinion and analyzing large-scale social media data.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Pak, A., Paroubek, P.: Investigating Twitter as a corpus for sentiment analysis and opinion mining. Language Resources and Evaluation Conference (2010)

[2] Go, A., Bhayani, R., Huang, L.: Utilizing distant supervision for Twitter sentiment classification. Stanford CS224N Project Report (2009)

[3] Thelwall, M., Buckley, K., Paltoglou, G.: Detecting sentiment strength in social web data. Journal of the American Society for Information Science and Technology 63(1), 163–173 (2012)

[4] Bollen, J., Mao, H., Zeng, X.: Predicting stock market trends using Twitter mood. Journal of Computational Science 2(1), 1–8 (2011)

[5] Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good, the bad and the OMG! ICWSM (2011)

[6] Barbieri, F., Bonafonte, A., Caputo, A.C.: Modeling tweet sentiment. Proceedings of the First Workshop on Unsupervised Learning in NLP, 53–63 (2012)

[7] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. Workshop on Languages in Social Media, 30–38 (2011)

[8] Bermingham, A., Smeaton, A.F.: Using Twitter to monitor political sentiment and predict elections. Workshop on Sentiment Analysis (2010)

[9] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with Twitter. ICWSM 10(1), 178–185 (2010)

[10] Poria, S., Cambria, E., Gelbukh, A.: Aspect extraction using deep convolutional neural networks. Knowledge-Based Systems 108, 42–49 (2016)

[11] Smailovi´c, J., Grˇcar, M., Lavraˇc, N., Zˇnidarˇsiˇc, M.: Active learning for sentiment

[12] analysis. Information Sciences 222, 17–29 (2013)

[13] Barbosa, L., Feng, J.: Robust sentiment detection on Twitter. COLING, 36–44 (2010)

[14] Mohammad, S.M., Turney, P.D.: Emotion lexicon using Mechanical Turk. NAACL Workshop, 26–34 (2010)

[15] Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis. LREC (2010)

[16] Saif, H., He, Y., Ferna´ndez, M., Alani, H.: Adapting sentiment lexicons. ACM TIST 5(3), 1–20 (2014)

[17] Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity. Computa- tional Linguistics 35(3), 399–433 (2009)

[18] Read, J., Carroll, J.: Domain-independent sentiment classification. CIKM Work- shop, 45–52 (2009)

[19] Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment insights. ICWSM (2011)

[20] Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment analysis in Twitter events. JASIST 63(1), 247–266 (2012)

[21] Gon¸calves, P., Arau´jo, M., Benevenuto, F., Cha, M.: Comparing sentiment analysis methods. ACM COSN, 27–38 (2013)