

Sentiment Analysis of Twitter Data Using Machine Learning Techniques

Mohammad Shaad¹, Neelesh Gupta², Kuldeep Mishra³, Himanshu Kumar⁴, Prof Ruchi Patel⁵,
Prof. Rajendra Arakh⁶

Department of Computer Science & Engineering
Shri Ram Institute of Technology, Jabalpur (M.P)

Abstract: In the age of social media, individuals regularly express their thoughts and emotions across various online platforms. Twitter, a prominent microblogging platform, serves as a prime example where users share their perspectives on diverse global events. Sentiment analysis, a crucial aspect of analyzing online discourse, involves discerning the emotional tone of text. This paper explores sentiment analysis on Twitter, employing machine learning and natural language processing techniques to categorize tweets based on their sentiment polarity. Various machine learning algorithms, including Vader, XGBoost, Random Forest, LSTM, and Bidirectional LSTM, are evaluated for their effectiveness in sentiment analysis. The study aims to assess the performance of these models in analyzing sentiments expressed on Twitter, with insights drawn from real-world data. Through sentiment analysis, organizations can gain valuable insights into public opinion, enabling informed decision-making across various domains.

Keywords: Crisis Management, LSTM, Sentimental Analysis, Tokenization, Vader

I. INTRODUCTION

Twitter has become a significant platform where people express strong emotions, making it a valuable resource for understanding sentiment. Sentiment analysis involves examining text to identify its underlying emotional tone. As social media platforms like Twitter gain prominence, sentiment analysis becomes crucial for businesses, organizations, and governments to grasp public opinion and make informed decisions [11]. Natural Language Processing (NLP) techniques play a vital role in sentiment analysis by enabling machines to understand and interpret human language [12]. These methods allow real-time analysis of tweets, offering insights into prevailing public sentiment trends [13]. Machine learning algorithms, a subset of NLP, learn from extensive datasets to accurately predict sentiment in new tweets. In this study, we aim to evaluate the effectiveness of machine learning systems in sentiment analysis on Twitter using NLP techniques. We will use a dataset containing tweets from the inaugural day of the "FIFA World Cup 2022" in Qatar. This dataset includes information such as tweet creation date, likes, source, content, and sentiment. After preprocessing to remove noise, we will employ machine learning methods such as Vader, XGBoost, Random Forest, and LSTM. To enhance accuracy, we will utilize count vectorizers and Gensim. These tools translate text into numerical data for machine interpretation. The efficacy of these

algorithms will be assessed based on various criteria like F1 score, accuracy, recall, and precision. By leveraging sentiment analysis, organizations can make informed decisions in areas like real-time monitoring, audience engagement, brand perception, fan experience improvement, and crisis management. Twitter sentiment analysis is vital for businesses to understand customer feedback and identify areas for improvement in products or services. It also aids in promptly addressing negative feedback online. Moreover, sentiment analysis can assist political campaigns in understanding public sentiment and adjusting messaging accordingly. During crises, sentiment analysis helps companies monitor social media and news channels for negative sentiments and take necessary actions. Additionally, marketers can use sentiment analysis to understand consumer preferences and tailor advertising campaigns accordingly [14-17].

II. LITERATURE REVIEW

This literature review delves into various studies exploring sentiment analysis on Twitter. Pak and Paroubek (2010) developed a classifier to categorize tweets as objective, positive, or negative, based on emoticons. Bahrawi (2019) employed Random Forest for sentiment analysis, achieving a 75% observation error margin. Shobana et al. (2019) investigated public sentiment through celebrity tweets, utilizing Python, Twitter API, and Text Blob. Kumar et al. (2020) attained high accuracy using Bi-LSTM for sentiment analysis. Jacob et al. (2021) employed clustering techniques to discern positive and negative sentiments in tweets. Cihan ÇILGIN et al. (2022) analyzed COVID-19-related tweets

using VADER and visualized sentiments using word clouds and N-grams. Lal Khan et al. (2022) proposed a CNN-LSTM model for sentiment analysis in English and Roman Urdu, achieving high accuracy rates. Dhanta et al. (2023) explored machine learning methods such as logistic regression and Naive Bayesian for sentiment analysis, with Naive Bayes exhibiting superior efficiency. These studies collectively demonstrate the effectiveness of various machine learning techniques in discerning sentiment from Twitter data, offering insights into public opinion and emotional trends on social media platforms.

III. METHODOLOGY

The sentiment analysis process has several steps. Firstly, we collect data and do some prep work. We got our dataset from Kaggle, containing tweets from the first day of the FIFA World Cup 2022 in Qatar. Then, we clean up the dataset by removing usernames, URLs, and common words that don't add much meaning (stopwords). After that, we standardize the words using lemmatization to make sure our analysis is consistent. To make our model more accurate, we add in some tools like CountVectorizer and Gensim Word2Vec Model. These help us turn the text into numbers, which the computer can understand better. We also use some techniques to pick out the most important features, so our model doesn't get overwhelmed with too much information. These techniques help prevent our model from becoming too complicated and taking too long to train. Once our data is ready, we use machine learning algorithms like Vader, XGBoost, Random Forest, and LSTM to figure out if the text is positive, negative, or neutral. These algorithms learn from examples in our dataset to make predictions about new tweets.

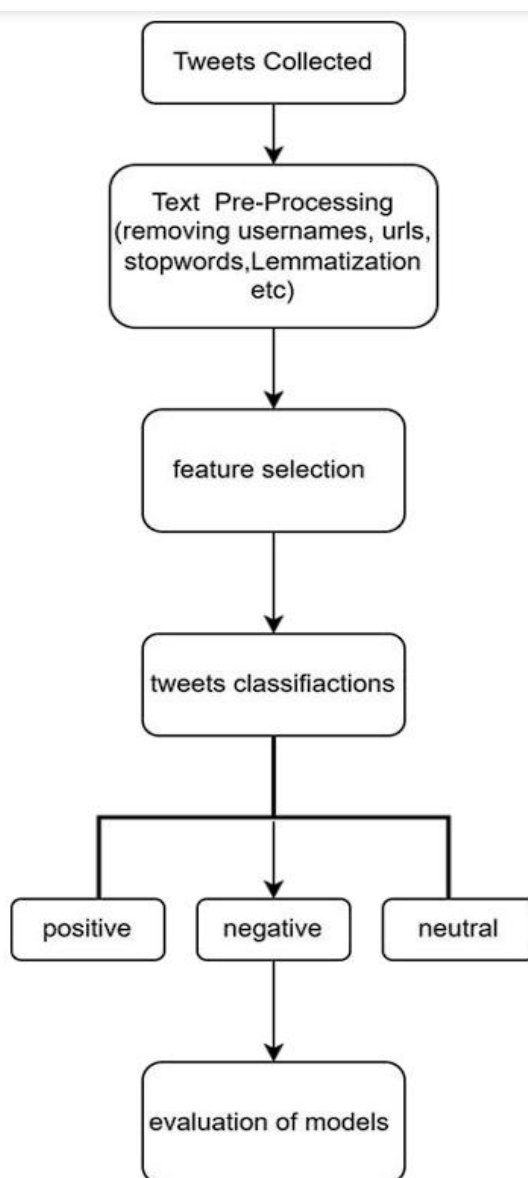


Figure 1: Methodology

A. Data Collection

We got this data from Kaggle, and it's all about tweets from the first day of the ‘‘FIFA World Cup 2022’’ in Qatar. [https://www.kaggle.com/code/aks777sp/fifa-world-cupday-1-tweets’]. It's got 22524 rows and 6 columns of information. These columns include the date the tweet was posted, how many likes it got, where it came from, the actual tweet text, and how people felt about it (the sentiment)

Unnamed: 0	Date Created	Number of Likes	Source of Tweet	Tweet	Sentiment
0	2022-11-20 23:59:21+00:00	4	Twitter Web App	What are we drinking today @TucanTribe ln@MadB...	neutral
1	2022-11-20 23:59:01+00:00	3	Twitter for iPhone	Amazing @CanadaSoccerEN #WorldCup2022 launch ...	positive
2	2022-11-20 23:58:41+00:00	1	Twitter for iPhone	Worth reading while watching #WorldCup2022 htt...	positive
3	2022-11-20 23:58:33+00:00	1	Twitter Web App	Golden Maknae shining brightlnhttps://t.co/...	positive
4	2022-11-20 23:58:28+00:00	0	Twitter for Android	If the BBC cares so much about human rights, h...	negative

Figure 2: The Dataset

B. Pre-Processing

To make our model work better, we need to get the data ready first. We start by cleaning up the tweets we collected. We take out things like usernames, web links, and common words that don't add much meaning. Once we've cleaned all that up, we break down the text into smaller parts called tokens. Then, we use a special tool called a lemmatizer to simplify the words to their basic form. This helps us analyze the text more effectively.

C. Machine Learning Models

a. VADER

In the nltk library for Python, there's a useful tool called VADER. It's made to understand text from social media, but it can work with other types of text too. VADER can figure out if a piece of text is positive or negative. It looks at a bunch of words that show different feelings to make its decision.

Word	Sentiment rating
Tragedy	-3.4
Rejoiced	2
Insane	-1.7
Disaster	-3.1
Great	3.1

Figure 3: Lexicon with valence rating

b. XGBoost

XGBoost is a type of machine learning that's part of a group called ensemble learning. It's really good at boosting, which means it takes a bunch of weak models and makes them stronger together. It mainly uses decision trees and some fancy tricks to make better predictions. XGBoost is known for being fast, analyzing important features, and handling missing data well. People use it for things like figuring out trends, sorting data, and making predictions about numbers or categories.

c. Random Forest

Think of a Random Forest like a team of people making decisions together. Each person represents a tree, and they all give their opinion to make better choices. This teamwork helps the model become stronger and more accurate. Random Forest is great because it's easy to use and can handle different types of problems, like sorting things into groups or predicting numbers. It's good at dealing with complex data without getting too caught up in the details, which makes it helpful for lots of different tasks in machine learning.

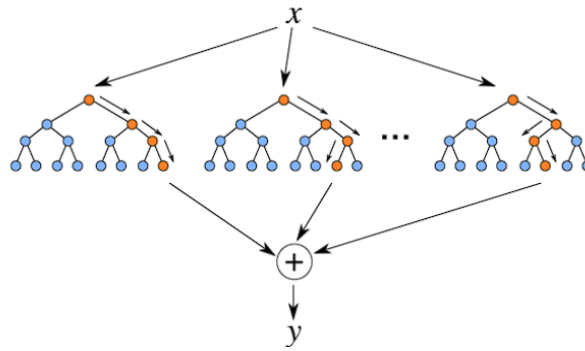


Figure 4: Random Forest

d. **LSTM** (Long Short-Term Memory)

LSTM (Long Short-Term Memory) is a type of RNN (Recurrent Neural Network) that can remember important information for a long time, even in a sequence of data. LSTMs are great for understanding and analyzing sequences, like text, sound, or time-based data. They're shown to work well for sentiment analysis on tweets because they can understand the context, handle different lengths of input, and recognize complex language patterns. They're useful for figuring out what people are feeling based on short and informal messages, which helps understand public opinions on social media. Bidirectional LSTM, or BiLSTM, is a special kind of sequence model with two layers of LSTM: one reads the sequence forward, and the other reads it backward. This setup is commonly used for tasks related to understanding language. The idea is that by looking at the input from both directions, the model can better understand how things are connected in the sequence.

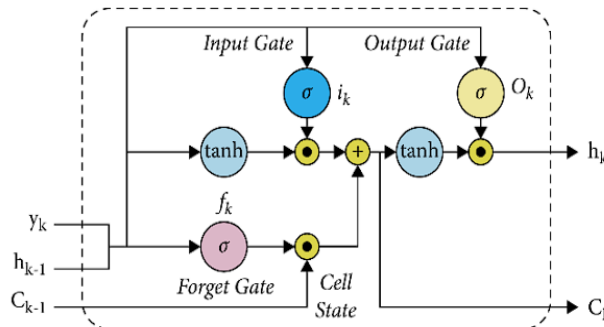


Figure 5: LSTM architecture

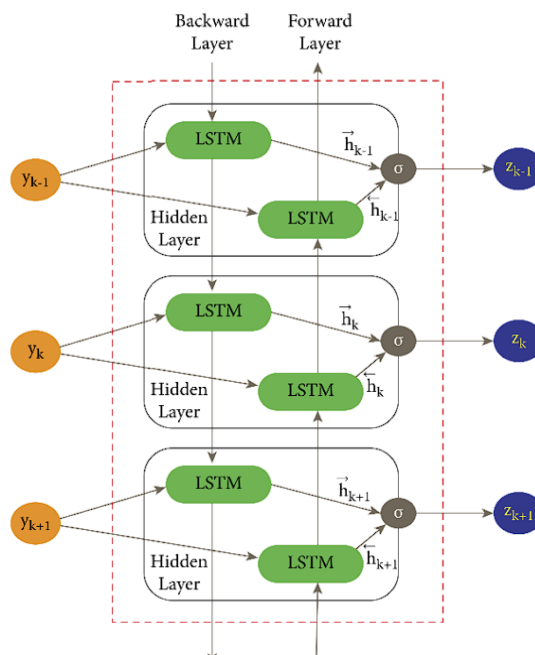


Figure 6: Bidirectional LSTM architecture

counts, and Figure 9 displays a Word Cloud representing sentiments. Figure 10 illustrates the trends of sentiment over time on the first day of the FIFA World Cup 2020 in Qatar. Emotion labels are counted at different time intervals to see how they change.

Various machine learning methods, including Vader, XGBoost with CountVectorizer, XGBoost with Gensim, Random Forest with CountVectorizer, Random Forest with Gensim, Single LSTM, and Bidirectional LSTM, are used for sentiment analysis. Among these, Bidirectional LSTM shows the best results, with an accuracy of 0.73. Compared to other models, Bidirectional LSTM performs the best. Additionally, Figure 11 displays the confusion matrix of the Bidirectional LSTM model, which has the highest accuracy. The F1-score, recall, and precision are all 0.73. Figure 13 shows the evaluation metrics of our machine learning methods.

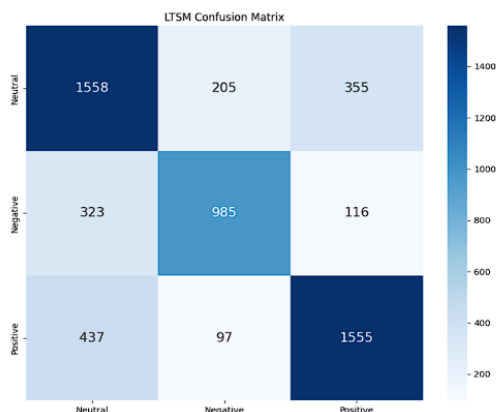


Figure 11: Confusion Matrix of Bi-LSTM Model

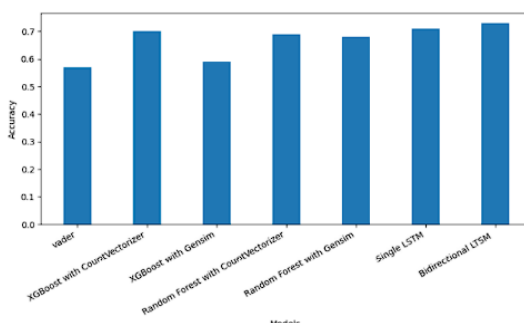


Figure 12: Accuracy scores of Machine learning Approaches

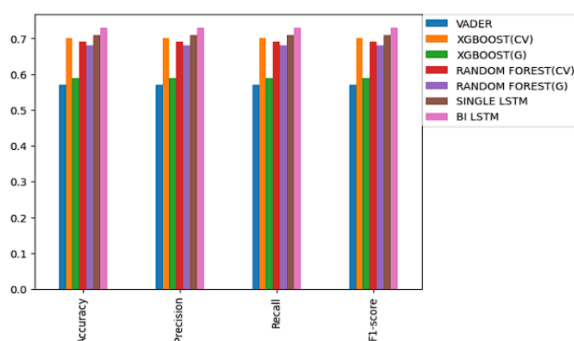


Figure 13: Graphical representation of evaluation performance of algorithms

V. APPLICATIONS

- Real-time monitoring:** It allows stakeholders to address negative opinions quickly while capitalizing on good sentiment.
- Audience Engagement:** Marketing teams may adapt their efforts based on sentiment analysis data to better connect with their target audience.
- Brand impression:** Sponsors can assess their brand's impression among fans and alter strategy appropriately.
- Fan Experience Enhancement:** Event organizers may use sentiment analysis to identify areas for improvement and then improve the entire fan experience, resulting in maximum satisfaction.

5. **Crisis Management:** Quickly identify possible controversies or unfavorable situations and take corrective steps to limit their impact.

VI. CONCLUSION

This study explores different machine learning methods like Vader, XGBoost with CountVectorizer, XGBoost with Gensim, Random Forest with CountVectorizer, Random Forest with Gensim, Single LSTM, and Bidirectional LSTM for analyzing sentiments on Twitter. Among these, Bidirectional LSTM stands out with the highest accuracy of 0.73. It's shown better performance compared to others. Initially, we load data from the Kaggle dataset "fifa_world_cup_2022_tweets," which contains tweets from the opening day of the FIFA World Cup in Qatar. Then, we clean the dataset by removing usernames, URLs, stopwords, and perform tasks like lemmatization and tokenization. We also create visualizations like sentiment distribution plots, word clouds showing positive and negative words, and time series sentiment trends during the event to get insights. Additionally, we calculate sentiment scores for each tweet using these models, covering aspects of positivity, negativity, and neutrality. Finally, we compare these models to understand their effectiveness. Looking ahead, Neural Network models hold potential to surpass other models in accuracy if properly optimized.

REFERENCES

- [1] Çilgin, C., Baş, M., Bilgehan, H. & Ünal, C. (2022). Twitter sentiment analysis during covid19 outbreak with VADER. *AJIT-e: Academic Journal of Information Technology*, 13(49), 72– 89. <https://doi.org/10.5824/ajite.2022.02.001.x>.
- [2] Dhanta, R., Sharma, H., Kumar, V. & Singh, H. O. (2023). Twitter sentimental analysis using machine learning. *International Journal of Communication and Information Technology*, 4(1), 71–83. DOI: 10.33545/2707661x.2023.v4.i1a.63.
- [3] Bahrawi, N. (2019). Sentiment analysis using random forest algorithm-online social media based. *Journal of Information Technology and Its Utilization*, 2(2), 29. <https://doi.org/10.30818/jitu.2.2.2695>.
- [4] Jacob, S. S. & Vijayakumar, R. (2021). Sentimental analysis over twitter data using clustering based machine learning algorithm. *Journal of Ambient Intelligence and Humanized Computing*. <https://doi.org/10.1007/s12652-020-02771-9>.
- [5] Ravi Kumar, G., Venkata Sheshanna, K. & Anjan Babu, G. (2021). Sentiment analysis for airline tweets utilizing machine learning techniques. In: *EAI/Springer Innovations in Communication and Computing*, pp. 791–799. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-49795-8_75.
- [6] Khan L, Amjad A, Afaq KM & Chang H-T. (2022). Deep sentiment analysis using CNNLSTM architecture of english and roman urdu text shared in social media. *Applied Sciences*, 12(5), 2694. <https://doi.org/10.3390/app12052694>.
- [7] Alexander Pak & Patrick Paroubek. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association.
- [8] Kumar, D. & Rao, S. (2020). A sentiment analysis of twitter data using bi-directional long short term memory. DOI: 10.1007/978-3-030-30271- 9_16.
- [9] Shobana, G., Vigneshwara, B. & Maniraj Sai, A. (2019). Twitter sentimental analysis. *International Journal of Recent Technology and Engineering*, 7(4), 343–346. <https://doi.org/10.46501/ijmtst061266>.
- [10] Dashrath Mahto, Subhash Chandra Yadav & Gotam Singh Lalotra. (2022). Sentiment prediction of textual data using hybrid convbidirectional-lstm model. *Mobile Information Systems*. <https://doi.org/10.1155/2022/1068554>.
- [11] I. Guellil & K. Boukhalfa. (2015). Social big data mining: A survey focused on opinion mining and sentiments analysis. *12th International Symposium on Programming and Systems (ISPS)*, Algiers, Algeria, pp. 1-10. DOI: 10.1109/ISPS.2015.7244976.
- [12] A. Świetlicka, D. Haczyk & M. Haczyk. (2023). Graph neural networks for natural language processing in human-robot interaction. *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Poland, pp. 89-94. DOI: 10.23919/SPA59660.2023.10274451.

- [13] K. S. Madhu, B. C. Reddy, C. Damarukanadhan, M. Polireddy & N. Ravinder. (2021). Real time sentimental analysis on twitter. 6 th International Conference on Inventive Computation Technologies, Coimbatore, India, pp. 1030-1034. DOI: 10.1109/ICICT50816.2021.9358772.
- [14] N. Deepa, J. S. Priya & T. Devi. (2023). Sentimental analysis recognition in customer review using Novel-CNN. International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1-4. doi: 10.1109/ICCCI56745.2023.10128627.
- [15] Y. E. Cakra & B. Distiawan Trisedya. (2015). Stock price prediction using linear regression based on sentiment analysis. International Conference on Advanced Computer Science and Information Systems (ICACISIS), Depok, Indonesia, pp. 147-154. DOI: 10.1109/ICACISIS.2015.7415179.
- [16] P. Khurana Batra, A. Saxena, Shruti & C. Goel. (2020). Election result prediction using twitter sentiments analysis. Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, pp. 182- 185. DOI: 10.1109/PDGC50313.2020.9315789.
- [17] A. Z. Adamov & E. Adali. (2016). Opinion mining and Sentiment Analysis for contextual online-advertisement. IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, pp. 1-3. DOI: 10.1109/ICAICT.2016.7991682.
- [18] Malde, Ravi. (2020). A short introduction to VADER. Towards Data Science. Available at: <https://towardsdatascience.com/an-shortintroduction-to-vader-3f3860208d53>.
- [19] Schott, Madison. (2019). Random forest algorithm for machine learning. Medium. <https://medium.com/capital-one-tech/randomforest-algorithm-for-machine-learningc4b2c8cc9feb>.