Sentiment Analysis of Cricket Tweets: A Comparative Study of Machine Learning and Deep Learning Approaches

Aditi Satsangi, Department of Electrical Engineering, Dayalbagh Educational Institute, Agra, India.

Abstract—**The advent of social media platforms has ushered in an era where vast amounts of data are generated daily. Twitter, as a microblogging platform, has become a rich source of information reflecting public sentiment, trending topics, and user engagement. This research delves into the multifaceted analysis of Twitter data, encompassing sentiment analysis, hashtag trends and machine learning applications. We are employing ensemble machine learning methods to enhance classification outputs in the field of sentiment analysis, which will increase the effectiveness and dependability of the suggested approach. Support Vector Machine, Random Forest, Logical Regression, and Deep Learning were used in the trials, and the findings show that the suggested model is provided better arrangement outcomes in terms of individual classifier accuracy.**

Keywords— SVM, Sentiment Analysis, Deep learning, Random Forest, Logistic Regression, Accuracy, Machine Learning, Hashtags, Accuracy.

I. INTRODUCTION

Twitter, a well-established web-based social networking platform, is renowned for its role in networking and microblogging, enabling users to send and receive messages known as tweets. As of 2021, Twitter boasts an impressive user base of 199 million monetizable daily active users, with India contributing significantly, accounting for 17.5 millions of these users.

In the contemporary era, social media has risen to prominence as a powerful influencer of public opinion [1]. Furthermore, it has become an active contributor to the enhancement of healthcare service delivery. Shifting towards a user-centric approach, social media allows for the rapid flow of information in near real-time, facilitating immediate interventions for individuals and communities in various healthcare settings, including hospitals, clinics, and homes.

Conversely, in the context of the stock market, news stories play a pivotal role in shaping investors' judgments and confidence in stock values. Engaging in stock market research empowers investors to make well-informed and successful investment decisions.

Sentiment analysis, an integral facet of text mining, holds significant relevance in text categorization. Specifically, Twitter tweet analysis represents a specialized subfield within sentiment analysis. It involves the classification of tweets into distinct categories. Twitter serves as an intermediary for individuals to exchange their perspectives, thereby serving as a valuable source of information for the public.

On Twitter, individuals express their views regarding various products in the market, encompassing both endorsements and criticisms. The ever-evolving market landscape is notably influenced by trending products, resulting in a dynamic representation of diverse opinions. It is important to highlight that Twitter's influence extends beyond product endorsements; it holds sway in numerous domains. Frequently, individual tweets possess the potential to exert influence and reshape prevailing circumstances.

Twitter tweets play an indispensable role in providing a comprehensive overview of public opinion regarding services or products in the market. They offer valuable insights into the collective sentiment and perceptions of diverse individuals. Sentiment analysis, commonly referred to as opinion mining, is the process of systematically analyzing product reviews and online content to gauge the prevailing sentiment or attitude towards a particular service or product.

Sentiment analysis operates as a powerful tool for understanding public opinion on social media platforms, effectively categorizing text into positive, negative, or neutral sentiments. Machine learning, a subfield of artificial intelligence, finds one of its major applications in sentiment analysis. Within the realm of consumer sentiment analysis, the need arises to develop an Automated Machine Learning Sentiment Analysis Model. However, applying models to text tweets poses a unique challenge due to the presence of significant noise and diverse content.

This paper aims to address this challenge through the classification of Twitter tweets. It seeks to evaluate the accuracy of sentiment classification for both positive and negative tweets, employing various methodologies, including Support Vector Machine (SVM), Random Forest, Logistic Regression, and Deep learning among others. The findings from this study hold substantial significance, enabling swift assessments of public sentiment towards a product or any other subject matter, such as news articles. Such insights are invaluable for businesses, marketing endeavors, and sports companies, facilitating informed decision-making and enhancing strategic planning.

II. REVIEW OF PREVIOUS WORK

In recent decades, there has been a significant surge in research and development related to sentiment analysis. This surge has led to the exploration of various applications for sentiment analysis. One particularly challenging task within this domain is the analysis of sentiment in textual data, such as messages and tweets. The intricacies of sentiment analysis in such text-based formats make it a demanding and complex endeavor. To tackle the above problem various solutions have been proposed previously [8].

In a previous work [2], P. D. Turney introduced the PMI-IR algorithm, which was applied to categorize reviews spanning various domains such as movies, banks, travel destinations, and vehicles. The algorithm achieved an accuracy rate of 74%. However, it was observed that analyzing movie reviews posed challenges due to the presence of distracting and less informative terms.

In a separate study [3], conducted by S. Asur and B. A. Huberman, the focus was directed towards the domain of movies, particularly in predicting box office revenue. Their research aimed to unravel the dynamics of Twitter discussions leading up to a movie's release and their correlation with box office earnings, both during the opening weekend and subsequent weekends. To validate their predictions, the study's results were cross-referenced with data from the Hollywood Stock Exchanges to assess accuracy.

In a previous study [4], the author introduced an approach for extracting sentiment from social media content. The proposed system leveraged innovative machine learning techniques, including Naïve Bayes (NB), Support Vector Machine (SVM), and Maximum Entropy. The model was implemented using popular data visualization tools, namely WEKA.

In another research endeavor [5], a team of authors delved into the domain of cricket match outcome prediction using tweets gathered from the Twitter social media platform. Their investigation focused on data from the 2014 IPL matches and the 2015 World Cup matches. The study involved the extraction of linguistic features for the purpose of predicting three key aspects: (1) the number of fan followers, (2) the quantity of tweets, and (3) score predictions derived from classifying tweets into positive, negative, and neutral categories. Interestingly, the study revealed that the SVM technique outperformed others, achieving an accuracy rate of 75%

In their work, N. Azam et al. [7] introduced an approach for sentiment analysis of tweets. The methodology involved tokenizing tweets using the n-gram technique, a method of text analysis. Features were extracted using Latent Dirichlet Allocation, enabling the identification of underlying topics in the tweets. Subsequently, the tweets were represented using a vector space model.

Furthermore, to identify dense regions within the tweet data, the authors applied the Markov clustering method. In their concluding remarks, the authors suggested that a hybrid model could yield improved results in sentiment analysis

. Horakova and Marketa have introduced an innovative technique focused on gathering tweets from various social media platforms to offer valuable business intelligence insights. Their sentiment analysis tool is structured with a two-layer framework, comprising a data processing layer and a sentiment analysis layer. Within this framework, the data processing layer takes care of critical functions such as data collection and data mining [6]."

III. Data set description

In this work we are focusing only the Cricket related tweets. This dataset is used from Kaggle named as “Cricket Tweets” given by Gabriel Preda, Data Scientist at Endava, Bucharest, Bucharest, Romania. Dataset is collected by using the Twitter API which is publicly available. Its size is 51862, It has 16 attributes. Main attributes which I have used for model such as 'id', 'user_name', 'user_location', 'user_description', 'user_followers', 'date', 'text', 'hashtags', 'source', 'retweets', 'favorites'. It is based on the cricket played between *1st September 2021 and 11th December 2021.*

IV. Methodology

To commence our analysis, we will categorize tweets into distinct groups based on their content. To accomplish this task, we will employ a diverse range of algorithms, including Support Vector Machine (SVM), Logistic Regression, Random Forest, and Deep learning. The tweets, initially represented as text strings, will undergo transformation into a numerical format through the utilization of the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This approach ensures compatibility with our chosen machine learning algorithms, allowing for comprehensive analysis.

A. Logistic Regression

In the realm of classification tasks, logistic regression stands out as a supervised machine learning technique. Supervised learning algorithms, including logistic regression, rely on labeled datasets that come equipped with an answer key. This answer key serves as the foundation for training and evaluating the model's accuracy.

In our pursuit of developing a sentiment classifier using logistic regression, we initiated the model training process with a sample dataset derived from Twitter [10]. The dataset, in its original form, comprises human-generated text, which can be particularly challenging for machine learning models to decipher. To facilitate the model's understanding of the text, we undertook a series of data pre-processing and cleaning steps.

The stages involved in pre-processing the tweets are as follows:

i. **Lower Case:** The tweets were uniformly converted to lowercase, ensuring consistent text representation.

ii. **HashTags:** Recognizing the valuable information contained in hashtags, we retained them for analysis, replacing the term 'hashtag' with the hash symbol.

iii. **Detaching Twitter Handles (@userid):** Twitter handles were separated from the text to enhance clarity.

iv. **Removing Special Characters, Numbers, and Punctuations:** Extraneous characters, numeric values, and punctuation marks were systematically eliminated.

v. **Tokenization:** The tweets underwent tokenization, a process involving the segmentation of text into its most fundamental and meaningful components

Model Building: For constructing our sentiment classifier, we employed logistic regression. This method employs a logarithmic function to fit the data and predict the likelihood of a particular event occurring. To supplement your understanding and for use in your paper, it's important to note that the dataset utilized for logistic regression comprises a collection of Twitter data related to sentiment analysis. This dataset is a representative sample of tweets with various sentiment labels (positive, negative, neutral). The dataset was pre-processed as outlined above to ensure its compatibility with logistic regression modeling techniques. The logistic regression model was trained and evaluated on this prepared dataset, yielding insights into sentiment classification.

B. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a powerful machine learning algorithm with the primary objective of predicting a hyperplane within an N-dimensional space, where N represents the number of features in the dataset. SVM is rooted in computational learning theory, employing methodologies aimed at addressing complex structural problems.

At its core, the concept of a support vector machine revolves around the discovery of a decision boundary that maximizes the separation between two distinct classes within the data. These boundary-defining vectors are aptly termed 'support vectors'. It's essential to note that altering the position of the hyperplane becomes necessary when manipulating or removing support vectors. These fundamental principles serve as the building blocks for the Support Vector Machine algorithm.

C. *Random Forest Classifier*

The random forest classifier is used for solving classification and regression problems. It is one of the popular machine learning algorithms due to simplicity and the ease of implementation.

The versatility of the Random Forest algorithm extends its applicability into various facets of our daily lives. It finds utility as a feature selector, aids in building recommender systems, and excels in image classification tasks. Beyond these practical applications, its real-life impact encompasses critical domains such as fraud detection, the classification of loan applications, and disease prediction. It forms the foundational basis for the Boruta algorithm, an invaluable tool for identifying and selecting vital features within complex datasets.

In the context of our Twitter sentiment analysis research, the Random Forest algorithm proves to be an indispensable tool, offering the potential to enhance the accuracy and reliability of sentiment classification in the dynamic and nuanced landscape of social media.

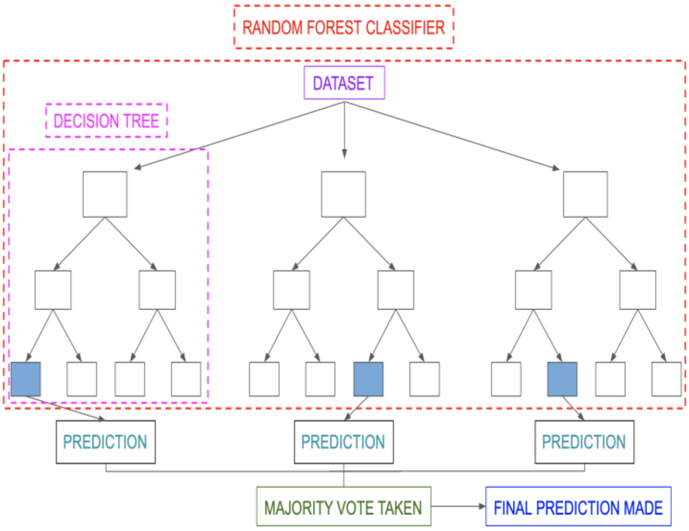


Fig.

Random Forest Classifier

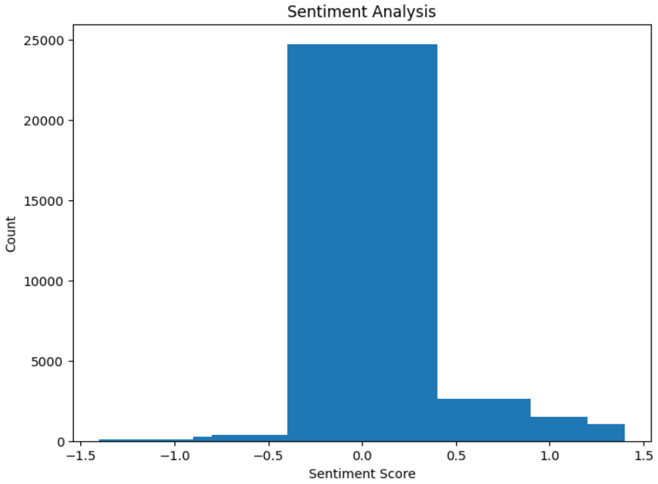
D. *Deep Learning*

Deep Learning, a subset of machine learning, has emerged as a dominant force in natural language processing tasks, including text classification. In this approach, we employ the TensorFlow and Keras libraries to preprocess and model textual data. The process starts with data preprocessing, involving tokenization and padding to prepare the text for deep learning. Next, we construct a sequential neural network model, incorporating essential layers like Embedding and LSTM, specialized in capturing semantic information and sequential dependencies within the text. The model is optimized for binary classification and monitored for accuracy during training. Deep Learning techniques offer remarkable capabilities for automatic feature extraction and understanding complex textual relationships, making them a go-to choose for tasks like sentiment analysis and text classification.

E, *Feature Extraction*

In the domain of sentiment analysis applied to cricket data, feature extraction plays a pivotal role in deciphering the emotions and sentiments embedded within the vast amount of textual information [9]. Feature extraction is the process of identifying and selecting relevant attributes or characteristics from the raw data that are most informative for sentiment classification.

Sentiment Scores: Calculating sentiment scores, which could be polarity scores (positive, negative) or emotion-specific scores (e.g., happiness, sadness) using sentiment analysis tools like TextBlob.



Mainly python is used for developing algorithms. Some packages used are as follows

Pandas:

Pandas is python package. It has functions for analyzing, cleaning, exploring, and manipulating data.

Numpy:

Numpy is a python library which is used for working with the matrices.

Matlplotlib:

Matplotlib is a low level graph plotting library in python that serves as a visualization utility.

Wordcloud:

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.

V. Results

To strengthen the credibility of our proposed sentiment analysis approach for cricket-related Twitter data, we have conducted a comparative analysis. This analysis involves assessing the performance of our method against other established sentiment analysis algorithms.[11]

A. *Logistic Regression*

Logistic Regression implementation is carried out from the dataset collected from Kaggle. The dataset collected about 51862 tweets from the Twitter application. The method's accuracy was estimated as 98.0%.

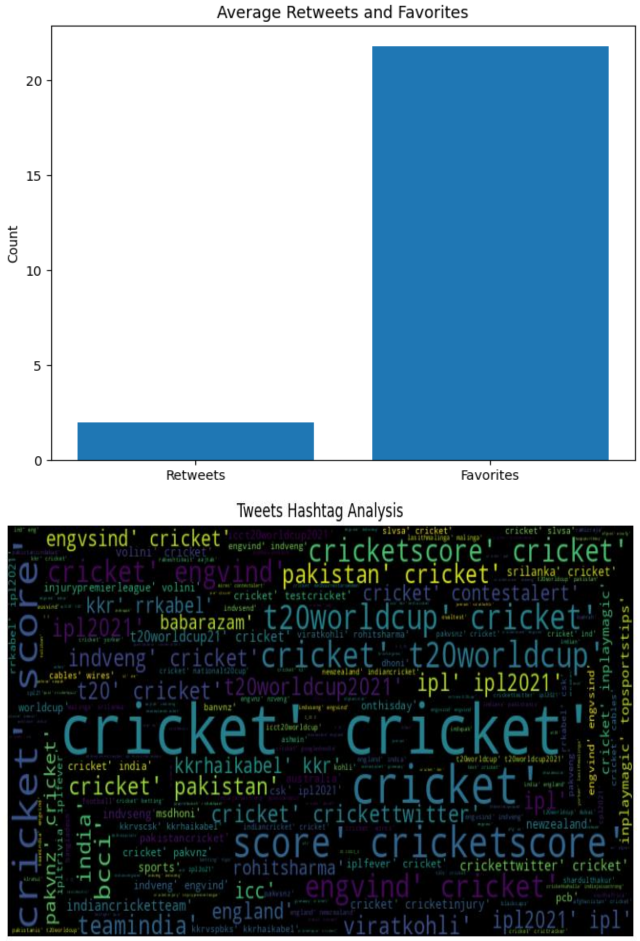


Fig. Average Retweets and Favorites Counts and representation

B. *Support Vector Machine*

The accuracy for Support vector machine (SVM) is 98.1%. This provides the highest accuracy in comparison to other machine learning algorithms like Random Forest tree and logistic regression for the given dataset.

C. *Random Forest Classifier*

The implementation of the Random Forest Classifier is carried out on a dataset consisting of around 25000 tweets. The accuracy for the random forest classifier is 65.9%. The report consisting of accuracy, f- score and precision is given below.

D. *Deep Learning*

The accuracy for the model developed using the TensorFlow, CNN gives the accuracy as 94.2% which is more accurate than the random forest classifier. Our CNN model's superior performance underscores the effectiveness of deep learning in handling complex sentiment analysis tasks, especially when dealing with unstructured data like tweets. CNN's ability to automatically learn and extract relevant features from textual data contributed to its impressive accuracy. The summary of the CNN model is given below.

TABLE 1: PERFORMANCE OF ALGORITHMS

Sequence No.	Algorithm Used	Accuracy %
1.	Logistic Regression	98.0%
2.	SVM	98.1%
3.	Random Forest	65.9%
4.	Deep learning	94.2%

TABLE 2: SUMMARY OF DEEP LEARNING MODEL (SEQUENTIAL)

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 100, 128)	640000
lstm (LSTM)	(None, 128)	131584
dense (Dense)	(None, 1)	129

TABLE 3: CLASSIFICATION REPORT FOR RANDOM FOREST

	precision	recall	f1-score	support
Positives	0.77	0.85	0.81	396
accuracy			0.66	519
macro avg	0.04	0.04	0.04	519
weighted avg	0.60	0.66	0.63	519

VI. CONCLUSION

In conclusion, the increasing reliance on internet evaluations necessitates the collection of diverse perspectives to facilitate learning from others' viewpoints. Sentiment analysis can be applied to various types of review data, making it a versatile tool. This paper contributes to the field by evaluating the performance of multiple sentiment classification algorithms and proposing a method to enhance classification accuracy. These studies compare various classification methods and select the most accurate approach for sentiment classification.

Traditional approaches indicate that the SVM classifier outperforms the other algorithms in terms of accuracy.

VII. FUTURE DIRECTIONS

In this paper, we utilized various machine learning and deep learning algorithms for predictions. While we applied Convolutional Neural Network (CNN) to our dataset, there is potential for further exploration of alternative algorithms like Recurrent Neural Network (RNN) and leveraging Natural Language Processing (NLP). These avenues represent promising directions for future research and improvement

VIII. REFERENCES

[1]. Prabha PM Surya, Lakshmi V Seetha and B Subbalakshmi, ―Analysis of user emotions and opinionsusing Multinomial Naiva Bayes classifier‖, IEEE, Coimbatore, 2019

[2] Peter D. Turney, DzThumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,dz Proceedings of the ИTth Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417- 424.

[3] Sitaram Asur and Bernardo A. Huberman,dzPredicting the Future With Social Media,dzarXiv:TTT¸.и□99vT [cs.CY] 29 Mar 2010

[4] Varsha Sahayak, Vijaya Shete, Apashabi Pathan,DzSentiment Analysis on Twitter Data,dz International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 1, Volume 2, January 2015, pp. 178-183.

[5] Raza Ul Mustafa, M. Saqib Nawaz, M. Ikram Ullah Lali, Tehseen Zia, Waqar Mehmood, DzPredicting The Cricket Match Outcome Using Crowd Opinions On SocialNetworks: A Comparative Study Of Machine Learning Methods,dzMalaysian Journal of Computer Science. Vol. 30(1), 2017, pp. 63-76

[6]. Diya Wang , Yixi Zhao, ―Using News to Predict Investor Sentiment: Based on SVM Model‖, 2019

[7] Nausheen Azam, Jahiruddin, Muhammad Abulaish, SMIEEE, and Nur Al-Hasan Haldar, DzTwitter Data Mining for Events Classification and Analysis,dz΄TTи Second International Conference on Soft Computing and Machine Intelligence, pp. 79-83.

[8] Agarwal, Amit, Bhumika Gupta, Gaurav Bhatt, and Ankush Mittal. ”Construction of a Semi-Automated model for FAQ Retrieval via Short Message Service.” In Proceedings of the 7th Forum for Information Retrieval Evaluation, pp. 35-38. ACM, 2015.

[9] Younus, Arjumand, et al. "TweetCric: A Twitter-based Accountability Mechanism for Cricket." *Web Engineering: 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings 17*. Springer International Publishing, 2017.

[10] Kannolli, Bharati S., and Prabhu R. Bevinmarad. "Analysis and prediction of sentiments for cricket tweets using hadoop." *International Research Journal of Engineering and Technology (IRJET)* 4.10 (2017).

[11] Siva Rama Rao, Akula VS, et al. "Sentiment Analysis: Twitter Tweets Classification Using Machine Learning Approaches." *Information and Communication Technology for Competitive Strategies (ICTCS 2021) ICT: Applications and Social Interfaces*. Singapore: Springer Nature Singapore, 2022.