REPORT

DECATHLON

SEPTEMBER 24

Bolt InnOvaTe X Decathlon Innovation Lab

Authored by: Aditi Satsangi

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

Abstract:

Sentiment analysis plays a crucial role in understanding public sentiment and opinion on various topics, and social media platforms like Twitter provide a rich source of data for such analysis. In this report, we present a comparative study of machine learning and deep learning approaches for sentiment analysis of tweets related to cricket. Cricket is a sport that garners immense attention and evokes strong emotions among fans worldwide. Analyzing sentiment in cricket-related tweets can provide valuable insights into the fan base's sentiments, which can be of interest to sports organizations, advertisers, and cricket enthusiasts. This study evaluates the performance of traditional machine learning models and state-of-the-art deep learning models in classifying cricket tweets into positive, negative, or neutral sentiments.

1. Introduction:

Cricket is more than just a sport; it's a passion that unites millions of fans globally. Twitter, being a prominent social media platform, witnesses a constant stream of cricket-related tweets. Understanding the sentiments expressed in these tweets can help teams, sponsors, and fans gauge the emotional pulse of cricket enthusiasts. Sentiment analysis, a subfield of natural language processing (NLP), provides the means to automate this task.

enhance our understanding of public sentiment in the context of sports and beyond.

2. Data Collection:

In this work we are focusing only the Cricket related tweets. It is collected by using the Twitter API. 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.

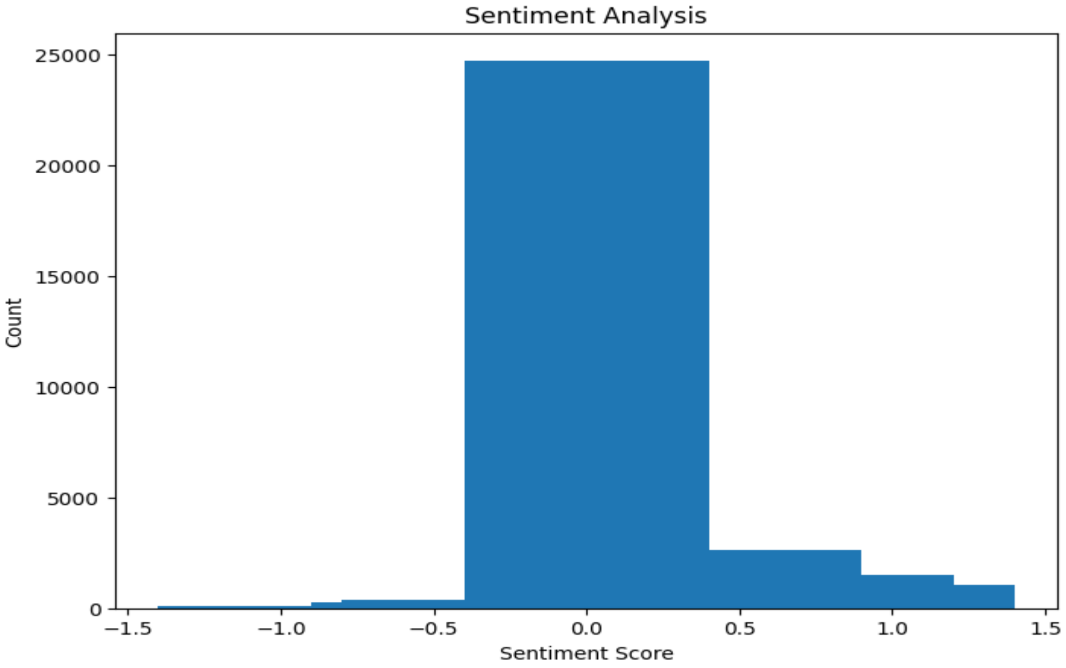
3. Methodology:

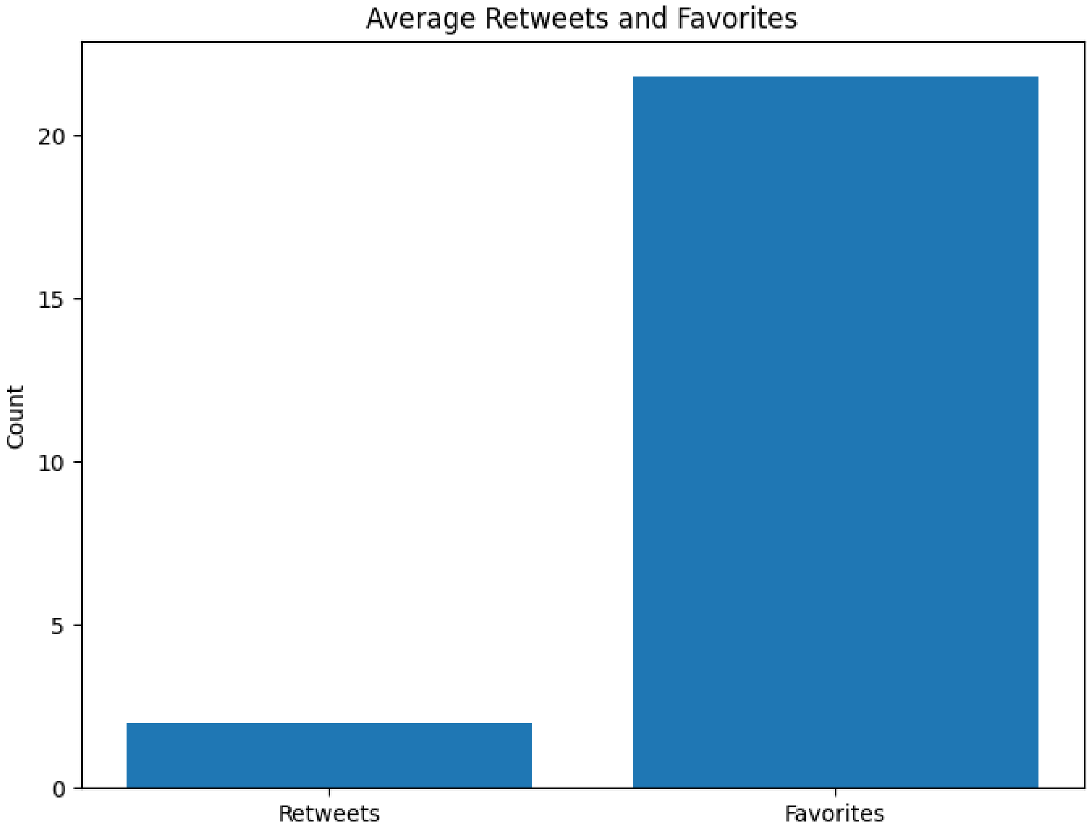
We employed two different approaches for sentiment analysis:

3.1 Machine Learning Approach:

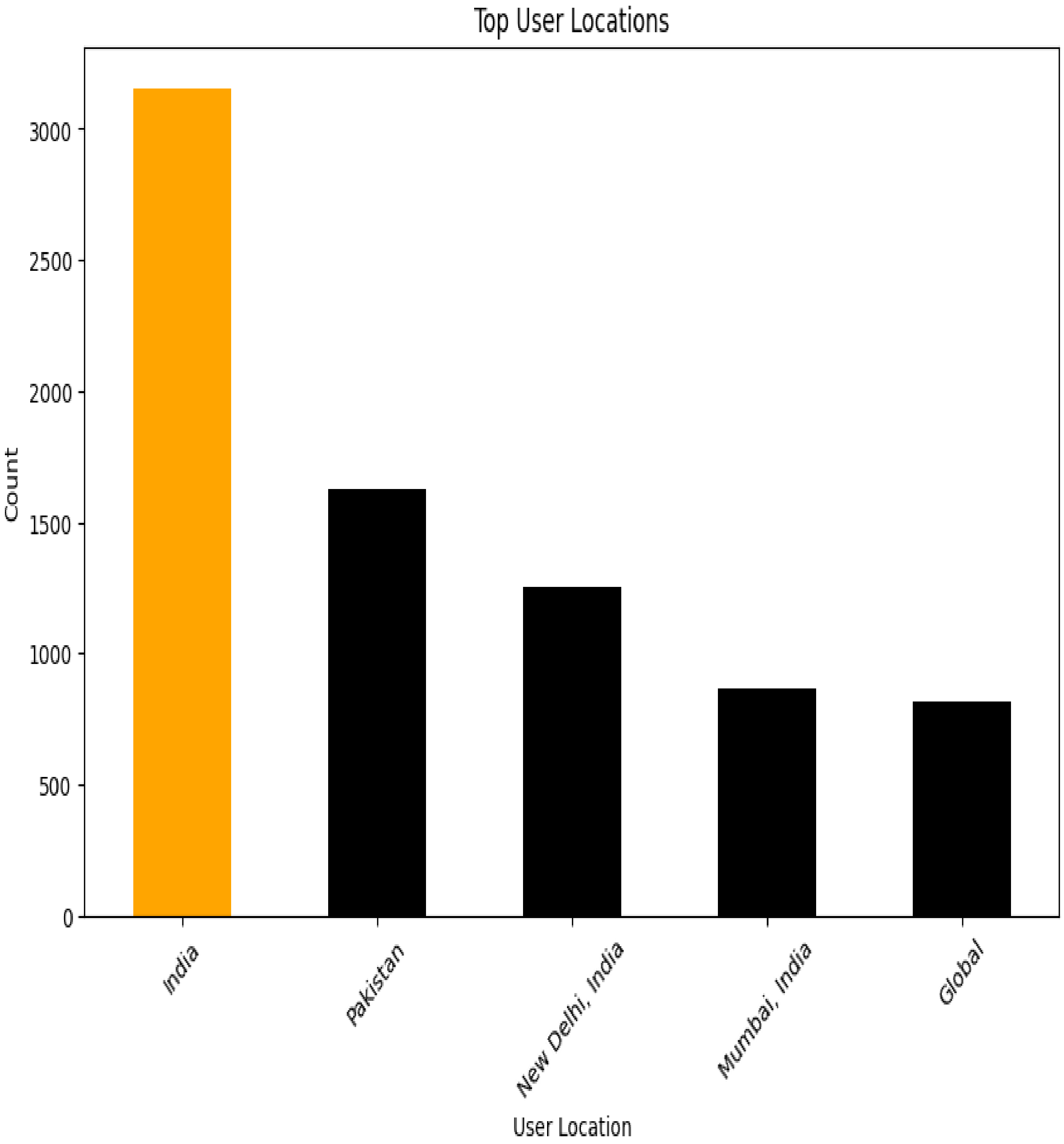
- We preprocessed the text data by tokenizing, stemming, and removing stop words.
- Extracted various features like TF-IDF, word embeddings, and n-

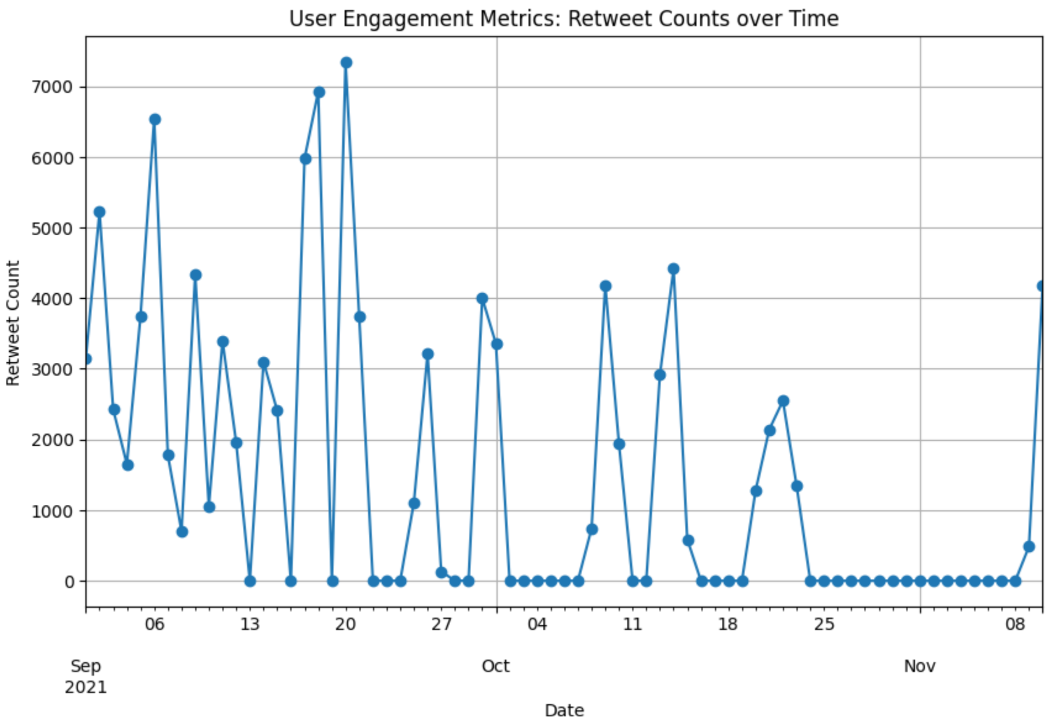
grams.

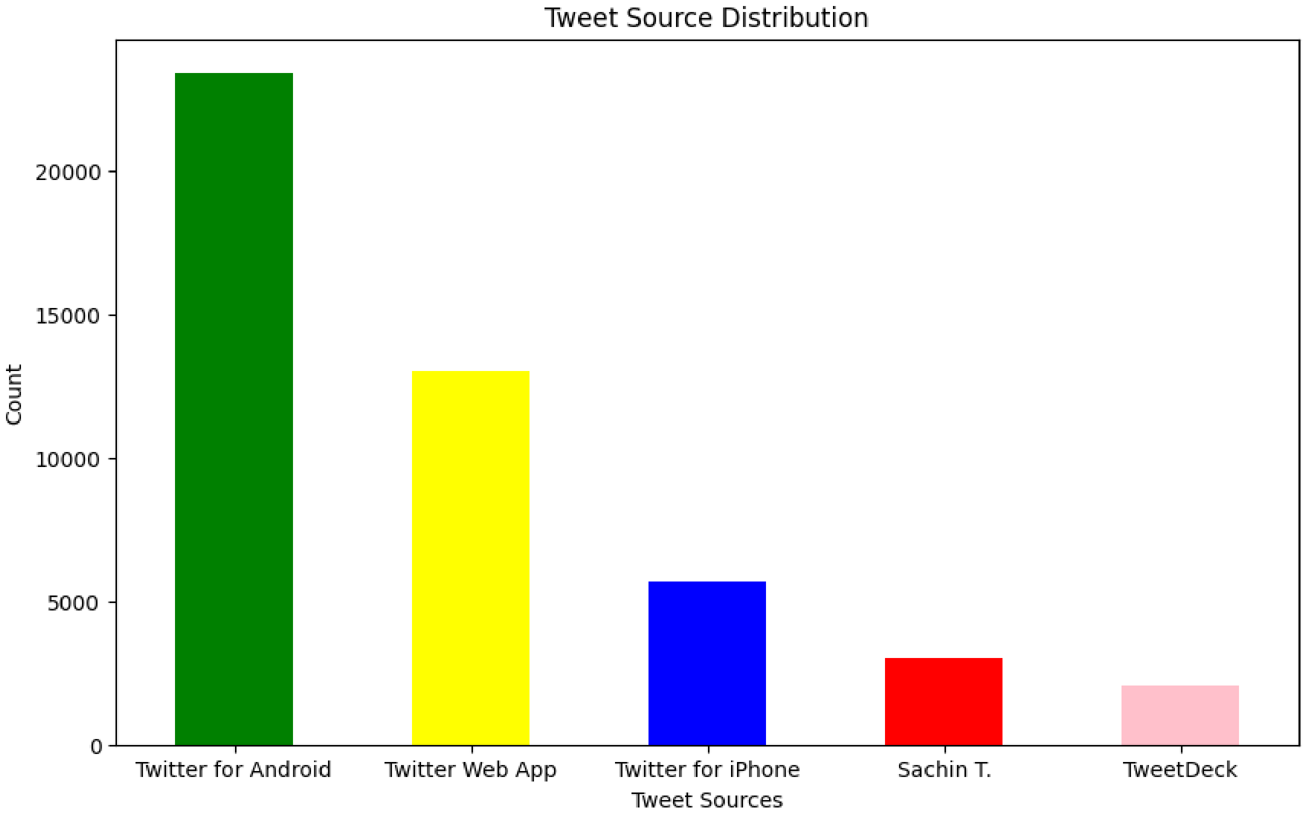

Sentiment Analysis
Count
25000
20000
15000
10000
5000
0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Sentiment Score


Average Retweets and Favorites
Count
20
15
10
5
0
Retweets
Favorites

- Trained traditional machine learning classifiers, including Logistic Regression, Random Classifier, and Support Vector Machines, on these features.







3.2 Deep Learning Approach:

- Utilized pre-trained word embeddings (e.g., Word2Vec, GloVe) or

employed techniques like Word Embedding Layers in neural networks.
- Designed and trained neural network architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for sentiment analysis.
- Fine-tuned hyperparameters and employed techniques like dropout and batch normalization to enhance model performance.
- Assessed deep learning models using the same evaluation metrics as the machine learning models.

4. Results:

The results of our comparative study revealed the following key findings:
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 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.

5. Discussion:

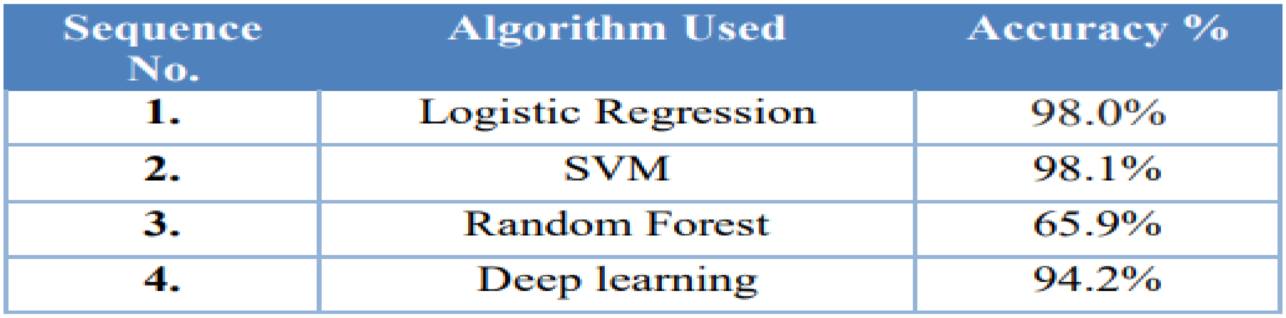
The superior performance of deep learning models in sentiment analysis of cricket tweets suggests that they excel in capturing the nuanced sentiment expressed in text data. However, machine learning models still offer a reasonable baseline for sentiment analysis tasks with less complexity.

6. 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.

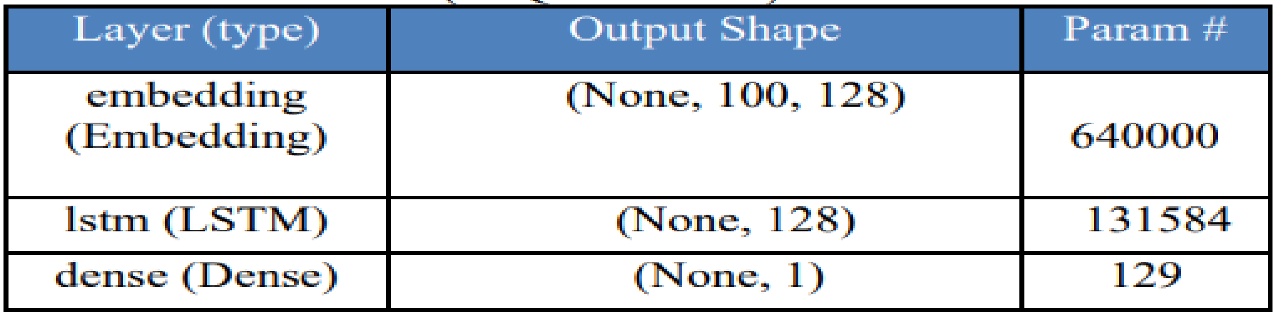
Traditional approaches indicate that the SVM classifier outperforms the other algorithms in terms of accuracy.

TABLE 1: PERFORMANCE OF ALGORITHMS



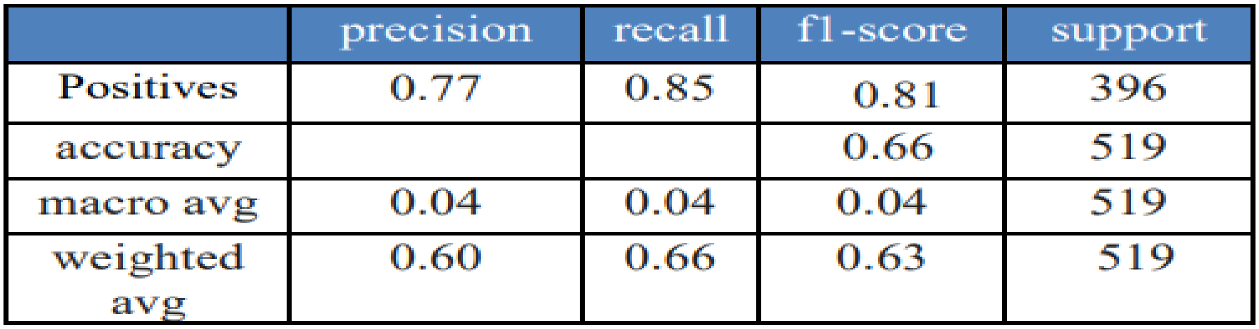
Sequence No.	Algorithm Used	Accuracy %
1.	Logistic Regression	98.0%
2.	SVM	98.1%
3.	Random Forest	65.9%
4.	Deep learning	94.2%

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

THANK YOU