Twitter Sentiment Analysis using Machine Learning Algorithms: A Case Study

Sheresh Zahoor
Electronics and Communication
Delhi Technological University
Delhi, India
sahrishzahoor@gmail.com

Rajesh Rohilla
Electronics and Communication
Delhi Technological University
Delhi, India
rajesh@dce.ac.in

***Abstract*— Sentiment analysis, also referred to as opinion mining or emotion extraction is the classification of emotions within a textual data. This technique has been widely used over the years in order to determine the sentiments, emotions within a particular textual data. Twitter is a social media platform that has been mostly used by people to express emotions for particular events. In this paper, we have collected tweets for a number of events, analyzed them using a number of Machine Learning algorithms like Naïve Bayes, SVM, Random Forest classifier and LSTM and compared the results.**

Keywords— Twitter Sentiment Analysis; Supervised Approach; Naïve Bayes; SVM; Random Forest Classifier; LSTM.

I. INTRODUCTION

Sentiment analysis or emotion AI refers to the use of natural language processing, computational linguistics to systematically extract emotions, sentiments, opinions i.e. the subjective information in a piece of textual data. Opinion mining has found its use mainly within the market research allowing a business to understand the sentiment regarding their products, services [1]. It not only allows the monitoring of opinions but also the likes and dislikes of people in general.

Over the years Twitter has developed as one of the leading social media platforms allowing people to express their views, opinions and sentiments with respect to an event, incident in the form of tweets. By monitoring these tweets, a company or a political party can easily understand how they are being perceived and what improvements could be made. It has changed the outlook to a great extent, not only providing all the necessary information but also making it possible for people to be open about their views. It has provided a platform to the common man so he can express his opinions and views. Thus, Social media could be considered as one of the best ways of monitoring the opinions of people when it comes to market research [2].

With the advances in Machine and Deep Learning, the ability to predict the sentiment within a textual data has increased. It is because of these algorithms that it has become possible to predict the sentiments with a great accuracy. Machine learning allows the learning new tasks without being explicitly trained or programmed. Sentiment analysis models can be used to predict not only the sentiment but other subjective information within a piece of text.

Several Machine learning and deep learning algorithms could be used for predicting the emotions, sentiments. In this paper we have used algorithms like Naïve Bayes, SVM, Random Forest Classifier and Long Short Term Memory Networks (LSTMs). This is a case study involving a number of events that have occurred in the last year like the Haryana Assembly Polls; a movie release The Sky is Pink; UNGA conference and a gathering Howdy Modi that occurred in Houston. The tweets related to these events were collected, fed to the algorithms and the results were compared.

II. LITERATURE REVIEW

It was Hearst [3] in 1992 and Kessler et al [4] in 1997 who initiated the classification of text based on the sentiments or emotions. There are two main approaches for classification of a piece of text, one is the lexical approach and the other is machine learning approach. Machine learning techniques gained interest because of the ability to extract features, capturing context [5]. These techniques are mostly used to detect the sentiment of an entire document and not just a few sentences.

In sentiment analysis, it is not always possible to detect the sentiment based on a single word, so the concept of n-gram extraction was introduced and found to be effective by Pederson in 2001 [6]. Other approaches involved selecting only a subset of the words detected using part-of-speech (POS) recognizer. Also, a step-wise approach of classification was developed later on which first involved the removal of objective sentences [1].

Over the years a lot of research has been done based on the tweets collected from twitter. It was Barbosa and Feng in 2010 who analyzed the Twitter data by determining the polarity of tweets on the basis of symbols, retweets, emoticons and even syntax of the Tweets. Initially, Naïve Bayes algorithm was used to analyze the sentiments of movie reviews [7]. It was Tong and Koller who later made use of Support Vector Machine in order to further improve the accuracy of prediction [8].

III. SENTIMENT ANALYSIS

With the increase in the social media platforms, there is an increase in the number of people expressing their sentiments and opinions, labelling these sentiments can be very useful for people who are looking forward to using these opinions in order to improve their products, services etc. [1]

Sentiment Analysis is the process of mining the text and determining the sentiment, opinion within the text. Sentiments can be determined using two approaches i.e. Unsupervised

978-1-7281-9785-2/20/$31.00 ©2020 IEEE

approach and Supervised approach or Machine Learning approach.

A. Unsupervised approach

It involves unlabeled dataset and the use of in-built libraries like TextBlob and VADER for predicting whether a piece of text is positive, negative or neutral.

B. Supervised approach

It involves labelled dataset and the use of Machine Learning algorithms like Naïve Bayes, SVM and so on.

1) Naïve Bayes- This is one of the supervised algorithm that is based on the probabilistic approach to classify the text to a particular class i.e. positive or negative [9]. This algorithm calculates the probability of all the words in the dataset and then classifies the tweets or text into particular categories. This algorithm is based on the Bayes rule.

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$$

Where

x, y = events

$P(x|y)$ = Probability of x given y is true

$P(y|x)$ = Probability of y given x is true

$P(x), P(y)$ = Independent probabilities of x and y

Steps involved:

- Splitting data into training and test set
- Building a vocabulary based on words present in the training set
- Matching the tweet content with the vocabulary
- Creating a feature vector
- Training the classifier using the feature vector i.e. training the model
- Testing the model using the test set

2) SVM- SVM is a supervised machine learning algorithm that has been used for both classification and regression problems. SVM classifies by determining a hyperplane to classify the data distributed in the n-dimensional space. The classification is done based on the mathematical functions called kernels and these kernels are used to determine a hyperplane. Two different classes exist on the opposite sides of the hyperplane and thus this plane could be considered a decision boundary which could help in simple classification of the data points available.
The equation of the hyperplane is as follows:

$$w.x - b = 0$$

Where

w = Weight vector

x = Input vector

b = Bias

The steps involved are:

- Gathering the training and test sets
- Vectorising the data
- Creating a SVM model for training
- Applying the model to the test set

3) Random Forest Classifier- This is a classifier that is composed of several Decision Trees as a single Decision Tree suffers from several disadvantages like noise which can affect the overall results and performance. However, Random Forest classifier has several advantages over the Decision Tree like it is robust when compared to a Decision Tree and the reason being, a Random Forest Classifier uses the concept of Bootstrapping i.e. each tree is trained on a different training data as we divide the training data into subsets equivalent to the number of trees. This ensures that each tree has a different and its own training data.

4) LSTM- Long Short Term Memory Networks which are mostly referred to as LSTMs are special kind of Recurrent Neural Networks that have the capability to learn long-term dependencies. These networks are widely used and over the years have improvised tremendously. Earlier it was RNN that was widely used to understand the sentences based on the meaning of previous words, these form loops and thus allowing the information to be stored. This stored information allows the network to understand the complete sentence and helps in prediction. However, there are times when more context is required to understand and more information needs to be stored. Thus as the gap increases, RNN becomes obsolete and is not able to understand and predict the meaning. LSTM solves this issue as it is considered an advanced or extended version of RNN that is able to learn long-term dependencies. It is made up of three gates-

- Input gate
- Forget gate
- Output gate

These gates are sigmoid functions with value occurring between 0 and 1 where 0 means the gate is blocking everything and 1 indicates that gate is allowing everything. Input gate depicts the information that needs to be stored in the cell state, forget gate depicts the information that is to be removed whereas the output gate provides activation for the final output.

IV. METHODOLOGY

Sentiment analysis using Machine Learning Algorithms involves a number of steps, most of them involving processing of the data captured from twitter. The steps involved are:
Data Collection
Data pre-processing
Part of Speech Tagging
Sentiment Analysis using Machine Learning Algorithms

A. Data Collection

This step involves extracting data from twitter in the form of tweets and for that it is very important to have an account

on twitter. Other than that, permission is needed from twitter for collecting tweets. After obtaining permission, we extracted tweets and collected them in .csv files.

B. Data pre-processing

Once the data is collected in the .csv files, it is very important to pre-process the information and remove all the unnecessary content. A number of pre-processing steps are involved, which are as follows:

1) Tokenization: This involves removal of URLs, hashtags, at-mentions and breaking of a string into a list of tokens. It is mainly done in order to count the occurrence of words.

2) N-grams Extraction: This involves grouping accompanying words together into phrases called n-grams. This is done in order to improve the quality of sentiment analysis.

3) Stemming: Here words are replaced by their stems or roots i.e. reading is replaced by read.

4) Stop words removal: It involves removal of prepositions, articles that have high occurrence but do not have any influence on the overall sentiment of the text.

C. Part-of-speech Tagging

This is the process of tagging each word in terms of its part of speech.

D. Sentiment analysis using Machine Learning Algorithm

In this approach structured dataset is applied to the Machine Learning Algorithms. In this paper, we have compared the results obtained from Naïve Bayes, SVM, Random Forest classifier and LSTM.

V. CASES

A. Haryana Assembly Polls

Haryana Assembly polls were one of the important events that occurred last year, a number of tweets were recorded with this hashtag. We collected around 16,000 tweets in the raw form, fed them to an in-built library and then fed them to Machine Learning Algorithms like Naïve Bayes, SVM and so on to obtain the results more accurately.

```
In [10]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/ActualNaiveBayes.py',
wdir='C:/Users/Sahrish Zahoor/Desktop/Project code')
0.9750994548401356
Naive Bayes Accuracy: 97.51
0.968

Naive Bayes Accuracy with the Test Set: 96.80
```

Fig. 1. Haryana Assembly Polls accuracy using Naive Bayes

```
In [7]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/svmm.py', wdir='C:/
Users/Sahrish Zahoor/Desktop/Project code')
[nltk_data] Downloading package stopwords to C:\Users\Sahrish
[nltk_data]     Zahoor\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
positive    4831
negative    1955
Name: sentiment, dtype: int64
0.9605893186003683
```

Fig. 2. Haryana Assembly Polls accuracy using SVM

```
In [116]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/Randomforestclassifier.py', wdir='C:/Users/Sahrish Zahoor/Desktop/
Project code')
Loading train and test data
Data loaded
Applying TF-IDF transformation
Training Random Forest Classifier
Predicting on train data
Training accuracy: 99.08157700818611%
Predicting on test data
Testing accuracy: 95.94992636229755
```

Fig. 3. Haryana Assembly Polls accuracy using Random Forest Classifier

```
In [115]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/lstmaccuracy.py', wdir='C:/
Users/Sahrish Zahoor/Desktop/Project code')
Loading train and test data
Data loaded
Tokenizing and padding data
Tokenizing and padding complete
Creating the LSTM model
Model: "sequential_14"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_13 (Embedding)     (None, 19, 128)           256000
_________________________________________________________________
spatial_dropout1d_13 (Spatia (None, 19, 128)           0
_________________________________________________________________
lstm_13 (LSTM)               (None, 256)               394240
_________________________________________________________________
dense_13 (Dense)             (None, 2)                 514
=================================================================
Total params: 650,754
Trainable params: 650,754
Non-trainable params: 0
_________________________________________________________________

C:\Users\Sahrish Zahoor\Anaconda3\envs\venv\lib\site-packages\tensorflow_core\python\framework\indexed_slices.py:433: UserWarning:
Converting sparse IndexedSlices to a dense Tensor of unknown shape. This may consume a large amount of memory.
  "Converting sparse IndexedSlices to a dense Tensor of unknown shape. "
Train on 4342 samples, validate on 1086 samples
Epoch 1/10
4342/4342 [==============================] - 14s 3ms/step - loss: 0.4671 - accuracy: 0.7672 - val_loss: 0.1853 - val_accuracy: 0.9171
Epoch 2/10
4342/4342 [==============================] - 12s 3ms/step - loss: 0.1464 - accuracy: 0.9454 - val_loss: 0.1416 - val_accuracy: 0.9355
Epoch 3/10
4342/4342 [==============================] - 9s 2ms/step - loss: 0.0872 - accuracy: 0.9689 - val_loss: 0.1245 - val_accuracy: 0.9466
Epoch 4/10
4342/4342 [==============================] - 10s 2ms/step - loss: 0.0572 - accuracy: 0.9800 - val_loss: 0.1294 - val_accuracy: 0.9494
Epoch 5/10
4342/4342 [==============================] - 11s 2ms/step - loss: 0.0432 - accuracy: 0.9855 - val_loss: 0.1271 - val_accuracy: 0.9484
Epoch 6/10
4342/4342 [==============================] - 10s 2ms/step - loss: 0.0331 - accuracy: 0.9885 - val_loss: 0.1498 - val_accuracy: 0.9512
Epoch 7/10
4342/4342 [==============================] - 10s 2ms/step - loss: 0.0240 - accuracy: 0.9922 - val_loss: 0.1665 - val_accuracy: 0.9503
Epoch 8/10
4342/4342 [==============================] - 10s 2ms/step - loss: 0.0158 - accuracy: 0.9940 - val_loss: 0.1389 - val_accuracy: 0.9558
Epoch 9/10
4342/4342 [==============================] - 10s 2ms/step - loss: 0.0089 - accuracy: 0.9982 - val_loss: 0.1912 - val_accuracy: 0.9576
Epoch 10/10
4342/4342 [==============================] - 11s 2ms/step - loss: 0.0089 - accuracy: 0.9975 - val_loss: 0.2400 - val_accuracy: 0.9558
Saving the LSTM model
1358/1358 [==============================] - 1s 997us/step
Test accuracy: 0.9528718590736389
```

Fig. 4. Haryana Assembly Polls accuracy using LSTM

Further, the tweets for BJP and ML Khattar were collected so as to determine the accuracy of the results.

```
In [11]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/ActualNaiveBayes.py',
wdir='C:/Users/Sahrish Zahoor/Desktop/Project code')
0.9646288209606987
Naive Bayes Accuracy: 96.46
0.96

Naive Bayes Accuracy with the Test Set: 96.00
```

```
In [12]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/ActualNaiveBayes.py',
wdir='C:/Users/Sahrish Zahoor/Desktop/Project code')
0.9786910197869102
Naive Bayes Accuracy: 97.87
0.984

Naive Bayes Accuracy with the Test Set: 98.40
```

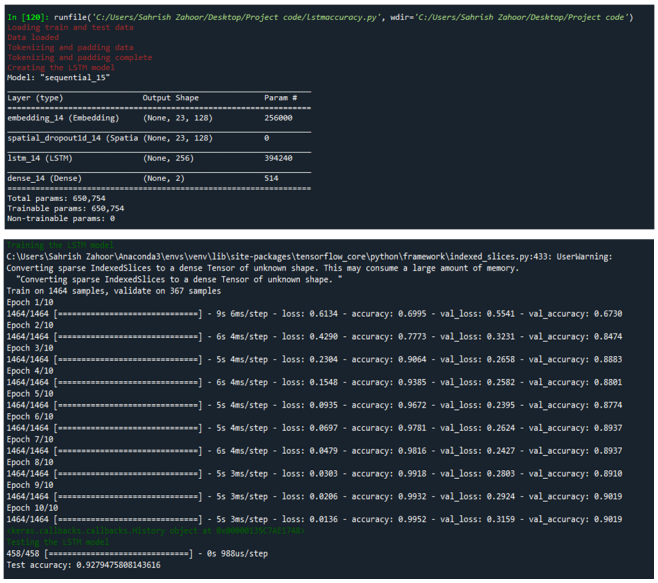
Fig. 5. BJP and ML Khattar accuracy using Naive Bayes

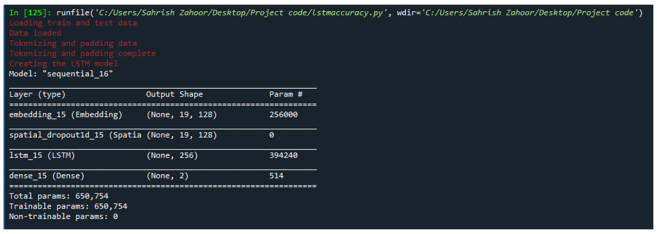
```
In [12]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/svmm.py', wdir='C:/
Users/Sahrish Zahoor/Desktop/Project code')
[nltk_data] Downloading package stopwords to C:\Users\Sahrish
[nltk_data]     Zahoor\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
positive    1612
negative     677
Name: sentiment, dtype: int64
0.8973799126637555

In [15]: runfile('C:/Users/Sahrish Zahoor/Desktop/Project code/svmm.py', wdir='C:/
Users/Sahrish Zahoor/Desktop/Project code')
[nltk_data] Downloading package stopwords to C:\Users\Sahrish
[nltk_data]     Zahoor\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
positive    529
negative    127
Name: sentiment, dtype: int64
0.9239543726235742
```

Fig.6. BJP and ML Khattar accuracy using SVM

Fig.7. BJP and ML Khattar accuracy using Random Forest





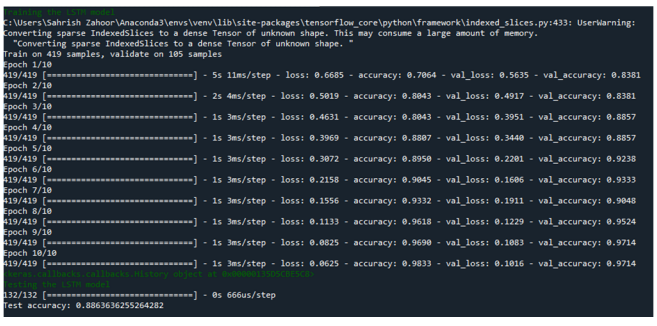


Fig.8. BJP and ML Khattar using LSTM

B. *The Sky is Pink*

This movie released on 11th October 2019 and was starring Priyanka Chopra, Zaira Waseem and Farhan Akhtar. It was received very positively by the critics and when fed to the in-built libraries, the response was around 62% friendly.

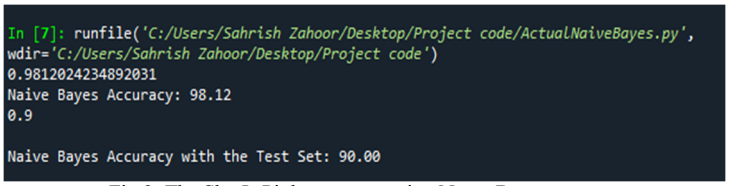


Fig.9. The Sky Is Pink accuracy using Naïve Bayes

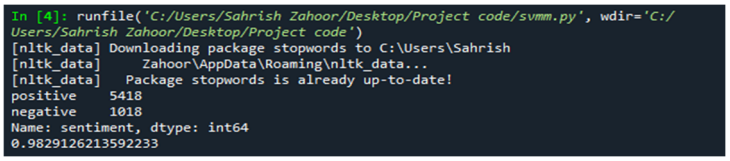


Fig.10. The Sky Is Pink using SVM

Fig. 11. The Sky Is Pink accuracy using Random Forest

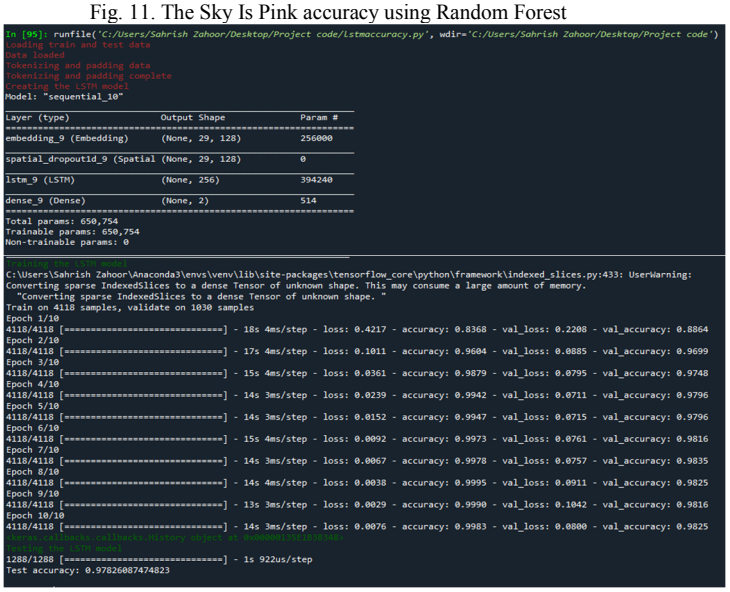


Fig.12. The Sky Is Pink using LSTM

C. *UNGA*

74th session of this assembly started on 17th September, a number issues were raised during this assembly, one among them being the climate change. Many such issues of dire importance were discussed by the world leaders so as to improve the living condition. Around 5000 tweets were collected and the accuracy of the sentiment was predicted based on a number of Machine Learning Algorithms.

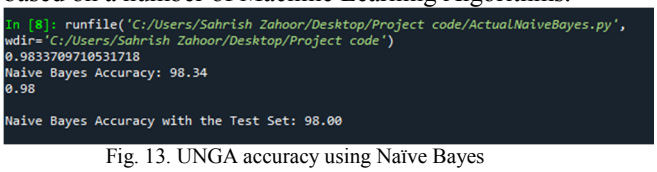


Fig. 13. UNGA accuracy using Naïve Bayes

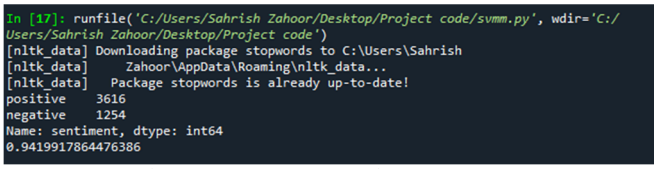


Fig. 14. UNGA accuracy using SVM

Fig. 15. UNGA accuracy using Random Forest Classifier

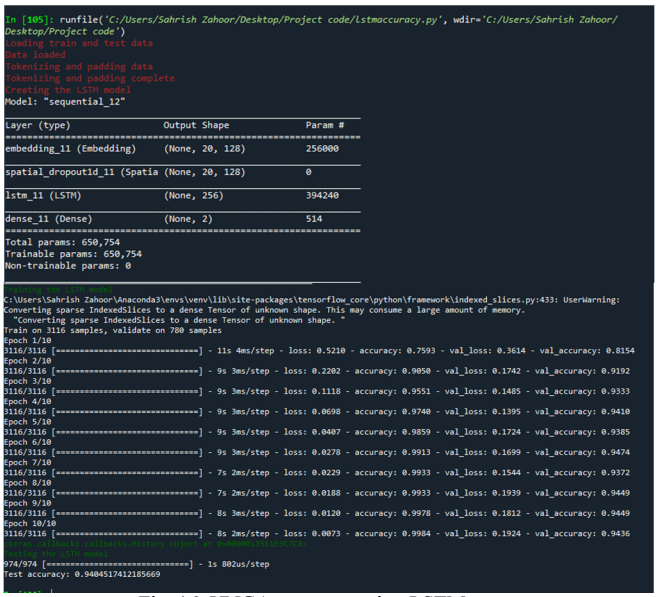


Fig. 16. UNGA accuracy using LSTM

D. *Howdy Modi*

This gathering occurred in Houston and was led by the Prime Minister of India Narendra Modi and the President of America Donald Trump.

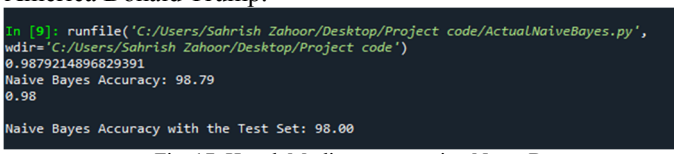


Fig. 17. HowdyModi accuracy using Naïve Bayes

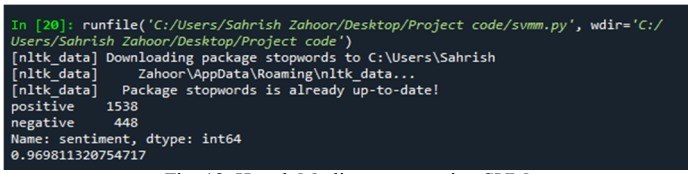


Fig. 18. HowdyModi accuracy using SVM

Fig. 19. HowdyModi accuracy using Random Forest Classifier

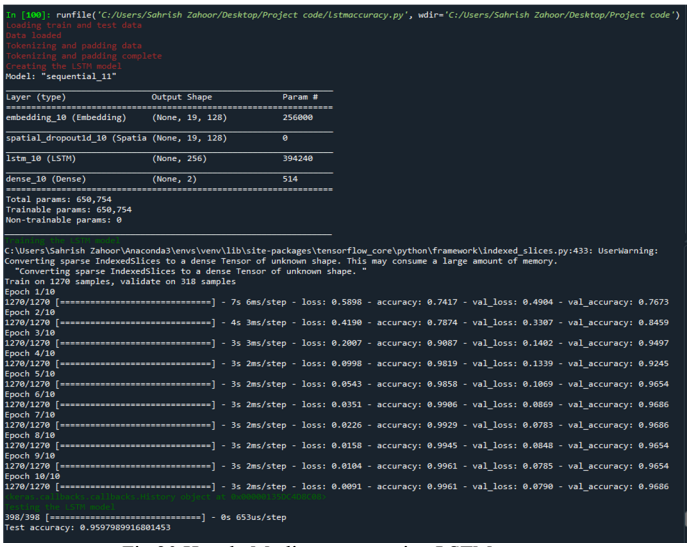


Fig.20 Howdy Modi accuracy using LSTM

VI. CONCLUSION

It can be concluded from this research that the sentiments can be predicted with more accuracy using Machine Learning and Deep Learning algorithms especially Naïve Bayes, SVM, Random Forest Classifier and LSTM. These have found their use in the field of Natural Language Processing especially in Sentiment analysis or to determine the subjective information like opinion, emotion within a piece of text. However, the accuracy with which each of them is able to predict the sentiment varies.

VII. RESULT

The final result obtained after feeding the different datasets to all the mentioned Machine Learning algorithms is mentioned in the Table I

TABLE I. SENTIMENT ANALYSIS

Machine Learning Algorithm	Haryana Assembly Polls	BJP	ML Khattar	The Sky is Pink	UNGA	Howdy Modi
Naïve Bayes	96.8%	96%	98.4%	90%	98%	98%
SVM	96.0%	89%	92.39%	98.29 %	94.1 9%	96.8%
Random Forest Classifier	95.94%	90.66 %	88.63%	98.2 %	94.6 6%	98.24 %
LSTM	95.28%	92.7 %	88.36%	97.8 %	94%	95.97 %

ACKNOWLEDGMENT

I would like to extend my gratitude to all the people who contributed to this research. I want to utilize this opportunity to thank my mentor Dr. Rajesh Rohilla for his immense support and encouragement. I want to thank my family, my brother who has always been my support, my friends especially Shams, who believed in my hard work and was my light throughout this project.

REFERENCES

[1] B. Pang , L. Lee, and S. Vaithyanathan, Thumbs up?:sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86.

[2] Edosomwan, Simeon, et al. "The history of social media and its impact on business." Journal of Applied Management and entrepreneurship 16.3 (2011): 79-91.

[3] Hearst M A (1992). Direction-based text interpretation as an information access refinement. In: Jacobs P, ed., Text-based intelligent systems: current research and practice in information extraction and retrieval. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, pp. 257–274.

[4] Kessler B, Nunberg G and Sch¨utze H (1997). Automatic detection of text genre. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Somerset, New Jersey, pp. 32–38.

[5] Polanyi L and Zaenen A (2006). Contextual valence shifters. In: Shanahan J, Qu Y and Wiebe J, eds., Computing attitude and affect in text: Theory and applications. Springer, pp. 1–10.

[6] Pedersen T (2001). A decision tree of bigrams is an accurate predictor of word sense. In: Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics. pp. 79–86.

[7] Boiy, E., Moens, M. A machine learning approach to sentiment analysis in multilingual Web texts. Inf Retrieval 12, 526–558 (2009).

[8] Tong S and Koller D (2002). Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2:45–66.

[9] Sentiment Analysis Using Naïve Bayes Classifier Kavya Suppala, Narasinga Rao IJITEE 2019