EasyChair Preprint

№ 8915

A Behavioural Analysis of US Election 2020 Using Deep Learning

Mohit Mathur, Ansh Agarwal and Shilpi Gupta

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

October 3, 2022

A Behavioural Analysis of US Election 2020 Using Deep Learning

Mohit Mathur
Department of Information Technology
JSS Academy of Technical Education
Noida, India
mohitmathur2600@gmail.com

Ansh Agarwal
Department of Information Technology
JSS Academy of Technical Education
Noida, India
anshagarwal592@gmail.com

Shilpi Gupta
Department of Information Technology
JSS Academy of Technical Education
Noida, India
shilpigupta@jssaten.ac.in

***Abstract*—Technology has evolved into a pivotal driving force for global communication. Twitter, Facebook, and Instagram are the most essential channels for expressing opinions on the daily developments that occur in and around the world. In this work, the tweets of US Election 2020 have been collected from twitter. After pre-processing the tweets, Sentimental Analysis using textblob and Behaviour Analysis using text2emotion are applied. The generation of feature vector is done using n-grams (unigram + bigram) for Machine Learning and neural networks keras for Deep Learning. Finally, the performance of the model is evaluated by using Deep Learning algorithms and then compared with Machine Learning algorithm. The model has achieved an accuracy of 84% using Bi-LSTM.**

***Keywords*—Machine Learning, Deep Learning, Sentiments Analysis, Behaviour Analysis, GloVe, Bi-LSTM.**

I. INTRODUCTION

Sentiment analysis (SA) has become a popular topic of study and research in recent years. The majority of social media data is controlled by Twitter [11]. On a daily basis, people discuss a wide range of topics, but it's impossible to gain insight merely by reading each of their viewpoints. Because humans struggle to make sense of large amounts of data, using an automated way will allow us to swiftly dig down into various consumer feedback groups that have been discussed on social media [7]. The process of predicting emotions in a word, sentence, or corpus of texts is defined as SA [1]. In this work, the extraction of user sentiment is identified using the textblob library in python. After extracting the sentiments, the work is also extended to Behavioural Analysis (BA) [2]. Human behaviour is extremely complex and cannot be summed up in a single number. Establishing a distinct emotional meaning for text can aid in the development of consumer relationships, motivation, and expectations for a brand or service. Simply categorising words into positive or negative categories is insufficient to appreciate the nuances of underlying tone, and polarity analysis provides minimal insight into the author's genuine meaning. The classification of a wide range of behaviours into states such as joy, fear, fury, surprise, and many others is known as behavioural detection. In this work, the US election tweets 2020 are collected and categorise tweets into positive, negative and neutral categories [2]. Political parties can use SA and BA to make decisions and to collect client feedback in a standardised style.

The performance of SA and BA can be analysed using different Machine Learning (ML) and Deep Learning (DL) approaches.

The objectives of the work are as follows:

- To apply Sentiment Analysis and Behaviour Analysis on US Election 2020 tweets.
- To evaluate the model using different ML and DL approaches.
- To compare the performance of different classifiers using performance metrics.

Section II covers the Related Work. The Proposed Methodology is discussed in Section III. Section IV covers the Implementation and Results. The Conclusion and Future Scope is discussed in Section V.

II. RELATED WORK

There is a lot of research work being done in the field of Sentimental Analysis and Behaviour Analysis. Some of the contribution is discussed in this section. The authors in the paper [12] demonstrated a textual analysis of Twitter data to determine public perception of the epidemic of anxiety, which was closely associated with the epidemic of coronavirus disease. In the paper [14], the authors had created their own handwritten neural network and used it to perform sentiment classification of tweets. They also compared the results with another neural network built with the popular machine learning library Keras. Achieve approximately 60% to 70% accuracy, depending on the parameters and input expressions used in the tweet. The authors of the paper [4] outline recent studies that have used deep learning to solve sentiment analysis problems and Emotional polarity. The model used term frequency inverse document frequency (TF-IDF) and word embedding on different datasets. A comparative study was also conducted on the experimental results obtained with various models and input characteristics.

In paper [3], the authors made an important contribution in this direction by analysing the behaviour of users belonging to both Facebook and Twitter. The author's research was based on data extracted from the web and can find important characteristics of these users related to privacy settings, friend choices, and activities. This was largely in line with the latest knowledge in this area. The authors of the paper [8] looked up tweets about the desperate purchase of toilet paper during the crisis. About half of the tweets had negative opinions. From this they concluded the findings affected how governments and key stakeholders used recent social media data to monitor and respond to public concerns. The authors in the paper [5] implemented a hybrid approach, which guarantees the sign of the total weight of the sentence according to its indirect meaning. The positive result is that opinions that were previously treated as neutral now make sense and inform their decisions. This hybrid approach used the concepts of a dictionary-based approach and a semantics-based approach. They used specific semantic rules to match words from a dictionary, assign their sentimental values, and analyse ironic or neutral tweets to get more information about their opinion.

III. PROPOSED METHODOLOGY

This section focuses on the proposed methodology work of the design of the system architecture. The work starts with the collection of data. Fig 3.1 shows the architecture of the proposed work.

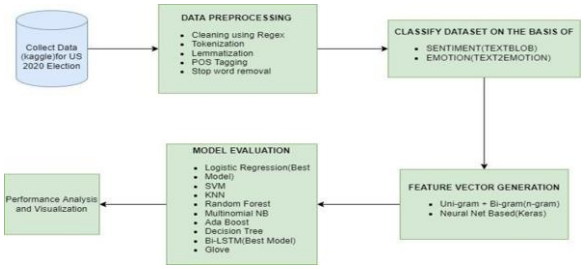

Fig 3.1 Architecture of the proposed work

3.1 Dataset Description

The data of the US 2020 Election is collected from Kaggle having a total of 1.8 million tweets with 22 columns such as ‘tweets’, ‘source’, ‘user name’, ‘candidate’ , ‘language’ etc It consists of data of different languages from different countries. We have used only English tweets which are around 1.1 million from which 650K is for candidate Trump and 450K is for candidate Biden.

3.2 Data Pre-processing

The pre-processing of the text data is an essential step as it makes the raw text ready for mining. The objective of this step is to clean noise those are less relevant to find the sentiment of tweets such as punctuation, emoticons, special characters, numbers etc.

Various pre-processing tasks such as cleaning, tokenization, lemmatization and POS tagging is applied on the dataset.

3.3 Classification

There are three main categories of tweets as positive, negative and neutral. For the classification of these tweets, we have used Textblob which is an opensource python library that helps to determine these sentiments of the user. It also helps us to identify the subjectivity and assign a particular score between 0 to 1.

For the extraction of emotions from the content, Text2emotion is used in this work. It is a Python package that helps to extract emotions from the content.

It processes all text messages and recognizes the emotions embedded in them and is compatible with 5 different emotion categories like happiness, anger, sadness, surprise and fear.

3.4 Feature vector generation

Feature vector is a n-dimension vector of numerical features corresponding to any n-dimensional feature with respect to this project work. Every tweet is represented in the form of n-dimensional feature vector for vocab. The generation of feature can be N-grams which is defined as continuous sequences of words, symbols, or tokens in a document. The value of N can be 1,2 and so on.

To generate N-grams, use the NLTK n-grams function with the parameter values as 1, 2 … N. However, we must first separate the text into tokens, which are then passed to the n-grams function.

3.5 Models Used

This section contains the description of various models such as Bi- Directional Long Short-Term Memory (Bi-LSTM) and Glove.

3.5.1 Bi-LSTM: Bi-LSTM (Bi-directional long short-term memory) Bidirectional long short-term memory (Bi-LSTM) is the most common way of developing a brain network that can store grouping data in the two bearings (future to past) and forward (ahead to future) (past to future). Bidirectional LSTMs contrast from traditional LSTMs in that the information streams in the two bearings.

3.5.2 GloVe: GloVe represents Global Vectors and is a disseminated word portrayal worldview. The model is a word vector portrayal procedure that is learned unaided. This is achieved by planning words into a significant space in which word distance is relative to semantic closeness.

Word embedding are performed for experimentation using pre-trained word vectors Glove. Glove version used 100 dimensions. 400,000 word GloVe embeddings computed on the 2014 English Wikipedia dump. Training is performed on an aggregated global word-word co-occurrence matrix that provides a vector space with meaningful substructures is mapped to the GloVe embedding. A co-occurrence count matrix is created for the above mapping. An embedding layer is created using the embedding matrix and set as the model's input layer. The model consists of 5 dense layers. The activation functions used were RELU and Sigmoid in the output layer. Hyper-parameters are fine-tuned to obtain the best model. After applying the different classifiers, the performance of the proposed model is measured in terms of Precision, Recall and F-1 score. In the next section, the Implementation details and Results are discussed.

IV. IMPLEMENTATION AND RESULT

In this work, the data is collected from having a total of 1.8 million tweets with 22 columns. We have used only English tweets which are around 1.1 million from which 650K is for candidate Trump and 450K is for candidate Biden. The model is implemented using Python.

The Glove model accuracy is represented in fig 4.1. At the start of the first epoch, the training accuracy is 0.55, and it achieves its maximum at the end of the last epoch. As the number of epochs increases, the data validation accuracy changes.

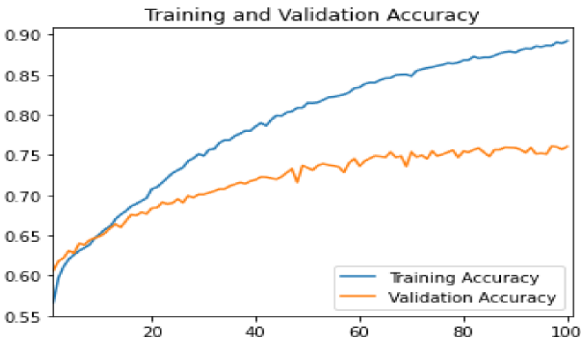

Fig 4.1 GloVe Deep Learning Model's Accuracy

As the number of epochs, or iterations, rises, the deep learning model's training loss lowers shown in fig 4.2 At the start of the first epoch, the training accuracy is highest , and at the end of the last epoch, it is lowest.

Fig 4.2 GloVe Deep Learning Model's Loss

The Bi-LSTM model accuracy is shown in fig 4.3 and 4.4 respectively. At the start of the first epoch, the training accuracy is 0.6, and it achieves its maximum at the end of the 100 epoch. As the number of epochs increases, the data validation accuracy changes.

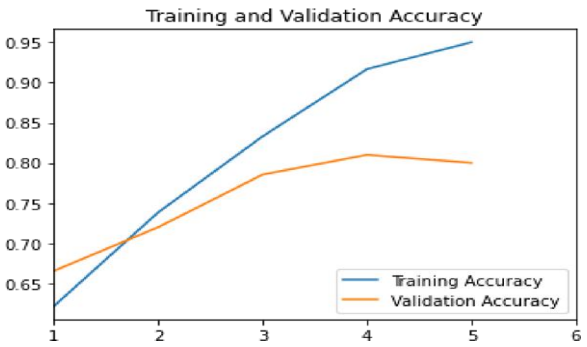

Fig 4.3 Bi-Directional LSTM Deep Learning Model's Accuracy

As the number of epochs, or iterations, rises, the deep learning model's training loss lowers shown in Fig 4.4. At the start of the first epoch, the training loss is highest, and at the end of the 10 epoch, it is lowest.

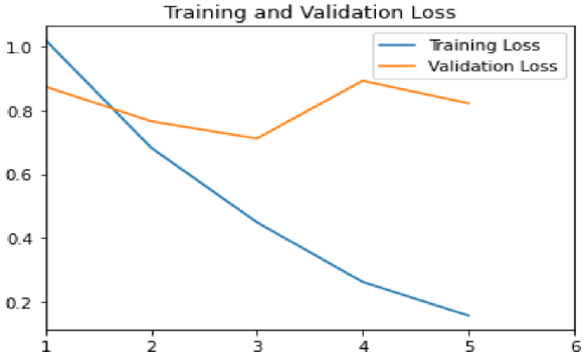

Fig 4.4 Bi-Directional LSTM Deep Learning Model's Loss

Table 1 depicts the performance of the model in terms of accuracy, precision, recall, and F1-score.

TABLE 1- Comparison of proposed model in terms of Accuracy, Precision and F1-score

Model	**Accuracy**	**Precision**	**F1-Score**
SVM	78.3	79.9	75.8
KNN	64.5	73.8	54.1
Ada Boost	72.3	71.9	69.1
Random Forest	74.3	77.6	70.7
Logistic Regression	84.2	84.4	83.1
Multinomial NB	73.1	72.2	71
Decision Tree	74.3	73.4	73.7
Bi-Directional LSTM	80	80.1	80
GloVe	75.7	74.1	74.5

When paired with unigram and bigram feature extraction over a restricted dataset of 20k tweets, Logistic Regression gives good results in terms of accuracy, precision and F1-score. The models achieved the maximum accuracy of 84.2% when compared with other models. Bidirectional LSTM performs well with Neural Net based Word Embedding using Keras and achieved the accuracy of 80%.

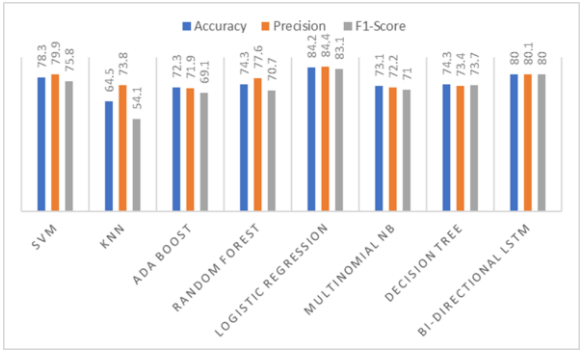

Fig 4.5 Bar graph of the different classifiers.

With a 75.8 f1-score, SVM surpasses Decision Tree in ML-based model evaluation, which is slightly higher

than that of Decision Tree (73.7). Among all the ML supervised learning models that have been deployed, Multinomial NB and Random Forest gave virtually comparable results of 71% F1-score value, which is higher than AdaBoost's F1-score.

V. CONCLUSION AND FUTURE SCOPE

In this work, the Bi-LSTM performance is better than Glove model. There is an increase of 5 % accuracy in this model. Out of the Machine Learning models, Logistic Regression achieved the best result. The scope of the work can be further extended by applying the behaviour analysis [13] model on other languages and emojis. The hashtags can also be used in future for prediction [15].

REFERENCES

[1] Alharbi, A. S. M., and de Doncker, E,. "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioural information". *Cognitive Systems Research*, *54*, 50-61, 2019.

[2] Balestrucci, A., De Nicola, R., Petrocchi, M., and Trubiani, C., "A behavioural analysis of credulous Twitter users". *Online Social Networks and Media*, *23*, 100133, 2021.

[3] Buccafurri, F., Lax, G., Nicolazzo, S., and Nocera, A., "Comparing Twitter and Facebook user behaviour: Privacy and other aspects". *Computers in Human Behaviour, 52*, 87-95, 2015.

[4] Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483.

[5] Gupta, S., Prashant, K., Sharma, L. K., and Panwar, M., "A Hybrid Method Proposed for Behavioural Analysis on Twitter Opinion Data Using Dictionary and Semantic based Approach". *International Journal of Advanced Research in Computer Science*, *9*(2), 2018.

[6] He, S., Wang, H., & Jiang, Z. H., "Identifying user behaviour on Twitter based on multi-scale entropy". In *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pp. 381-384, October, IEEE, 2014.

[7] Korakakis, M., Spyrou, E., and Mylonas, P., "A survey on political event analysis on Twitter". In *2017 12th international workshop on semantic and social media adaptation and personalization (SMAP)*, pp. 14-19, July, IEEE, 2017.

[8] Leung, J., Tisdale, C., Chung, J. Y. C., Chiu, V., Lim, C. C., and Chan, G., "Panic buying behaviour analysis of COVID-19 related toilet paper hoarding content on Twitter", 2021.

[9] Medhat, W., Hassan, A., and Korashy, H, "Sentiment analysis algorithms and applications: A survey". *Ain Shams engineering journal*, *5*(4), 1093-1113, 2014.

[10] Oliveira, J. E. M., Cotacallapa, M., Seron, W., dos Santos, R. D., and Quiles, M. G., "Sentiment and behaviour analysis of one controversial American individual on twitter", In *the International Conference on Neural Information Processing*, pp. 509-518, Springer, Cham, October, 2016.

[11] Qi, S., AlKulaib, L., and Broniatowski, D. A., "Detecting and characterising bot-like behaviour on Twitter". In *International conference on social computing, behavioural-cultural modelling and prediction and behaviour representation in modelling and simulation*, pp. 228-232. Springer, Cham, July, 2018.

[12] Raheja, S., and Asthana, A., "Sentimental analysis of twitter comments on COVID-19". In *2021 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence)*, pp. 704-708, January, 2021.

[13] Sharma, Parul, and Teng-Sheng Moh., "Prediction of Indian election using sentiment analysis on Hindi Twitter". *2016 IEEE international conference on big data (big data)*, IEEE, 2016.

[14] Sosa, P. M., and Sadigh, S., "Twitter sentiment analysis with neural networks". *Academia. Edu* , 2016. Sujay, R., Pujari, J., Bhat, V. S., & Dixit, A., "Timeline analysis of twitter users". *Procedia computer science*, *132*, 157-166, 2018.

[15] Sultana, M., Paul, P. P., and Gavrilova, M., "Identifying users from online interactions in Twitter". *In Transactions on Computational Science XXVI* pp. 111-124, Springer, Berlin, Heidelberg, 2016.