*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 10 Issue 6 June 2022

Hybrid Machine Learning Application in Classifying the Sentiment by Analyzing the Political Tweets

Authors: Jessica Kohli[(a)], Reetu Jain[(b)]

[(a)] Student, Class of 2023, Jamnabai Narsee International School, Narsee Monjee Bhavan 7, N S Road No. 7, JVPD Scheme, Vile Parle West, Mumbai, Maharashtra-400049, India

[(b)]Mentor, On My Own Technology Pvt. Ltd., 2A, Pace House, 7 Swastik Society, Gulmohar Road, Vile Parle West, Mumbai - 400049, India.

Email-id: reetu.jain@onmyowntechnology.com

Abstract:

In this present techno-socio world, social media has dramatically changed the lives of peoples. Social media, especially Twitter, is vastly used for sharing information and thoughts, expressing opinions, seeking support and many other such activities. Scientists and strategists try to analyse the sentiments of the common people from their tweets. By evaluating the sentiments of the common people, political scientists are capable of detecting early crises and the strategist could strategize the next move in favor of the political figures. Because of this reason the comprehensive intention of the present study is to develop a robust classification algorithm that can be applied for sentiment analysis from the political tweets. In order to achieve this objective a hybrid machine learning classification algorithm is proposed in the study. proposed hybridizes natural language processing (NLP) and long short term memory (LSTM). The NLP part of the model part of the proposed model extracts keywords that can rightly summarize the text of the tweets and correlate them with the positive or negative sentiments. On the other hand the LSTM part develops a predictive model that is capable of classifying the keywords into positive or negative sentiments. The potentiality of a proposed model lies in its applicability. Therefore the proposed model is applied to detect the sentiments of common people from their tweets. The Twitter data set comprises 1.6 million political tweets. The proposed model showed an accuracy of 78% and a loss of 0.456. The proposed model also showed a precision of 79% and F1 score of 0.78.

Keywords: Natural language processing, Long short-term memory, sentiment analysis, political tweets, Hybrid machine learning model

1. Introduction

Social media is evolving as a powerful platform for not only sharing information but also influencing opinions and decision making. In this present techno-socio world, social media has dramatically changed the lives of peoples. Social media is vastly used for sharing information and thoughts, expressing opinions, seeking support

and many other such activities [1]. Twitter is a microblogging social networking site that enables a user to share images, short videos and blogs with a limitation of 280-characters.

With the increasing number of internet-users, social networking sites especially twitter are gradually gaining popularity due to introduction of hashtags, usage limitation in characters and also tagging property [2]. With the increasing number of twitter users, it has become one of the major social sites to be used by the public figures in order to remain well-connected with their followers. Mostly, the political figures use twitter to remain connected with their apposition and fellow countrymen to laud their achievements, boost their morale, praise them, wish them and in certain tough scenarios appeal to them [3]. On the other hand, with the help of twitter , people can also show their likes, dislikes, respect, agony, support, oppose etc. to the political figures or to their policies, decisions and their speech.

In the present scenario, twitter has evolved as a platform where people can express their opinions and feelings towards their leaders. The researchers in this field evaluate the political based tweets from the people to predict their sentiments towards their leaders.

1.1. Literature review

The earliest known work that evaluates the political sentiment of people from microblogs was done in the paper [4]. In the study, the authors analyzed 104003 tweets to predict the result of German federal elections. The paper has set the foundation for predicting the political sentiments of people by analyzing the micro blogs. Traditionally, sentiment analysis is conducted by three different methods viz. Machine learning (ML), Lexicon based method and hybrid method [5].

Sentiment analysis is defined as the action of explaining the meaning of emotions such as positive, negative, or neutral through various text-mining methods and materials [6]. The ML techniques are mostly used to classify the sentiments behind the tweet into the classifications as identified in literature. In general, ML methods used for sentiment analysis includes natural language processing (NLP), neural network (NN), case based reasoning (CBR) and support vector machines (SVM). NLP methods aim at helping people to communicate with computers through the use of natural language. It extracts some keywords and classifies the sentiment of the message [7]. A typical NLP toolkit comprises tokenization, part-of-speech (PoS), tagging, chunking, named entity recognition and sentiment analysis [8]. In the reference [9], it is established that the NLP toolkits such as the NLTK [10], Stanford CoreNLP [11], and TwitterNLP [12] have tokenization, PoS tagging and NER modules in their pipelines. The second method for sentiment analysis from small messages includes the CBR technique. The techniques are designed to solve a new problem by remembering a previous, similar situation. Sentiment analysis can arguably be understood as a knowledge-based classification problem [13]. In the literature [14], the output obtained from the CBR is compared with the output from the rule based method of the sentiment analysis problem.

The third most common ML technique for sentiment analysis is the application of NN. It learns a function to understand and translate input data into the desired output. In the paper [15], the author has developed a deep learning NN model to analyze the movie reviews from the rottentomatoes.com website. The proposed NN model is improved in [16] and used for analyzing the movie reviews. A review of NNs for analyzing political sentiment

tweets can be found in [17]. The final most common ML technique for sentiment analysis is the SVM method. In these techniques, a hyperplane is discovered that can be used to categorize points from the dataset in a high dimensional space, thus leading to determination of the discriminant function that best separates the groups of data points. The paper [18] presents a systematic review of the literature that applies SVM for sentiment analysis.

Lexicon-based approaches, also known as symbolic approaches [19], use a sentiment dictionary to determine polarity, and sentiment scores are assigned to words to react to the positive, negative, or neutral attitude of the speaker [20]. There are three main schemes among lexicon based approaches namely, manual schemes, dictionary based schemes, and corpus-based schemes [21]. Finally, hybrid approaches are combinations of machine learning and lexicon-based techniques. More research papers are reviewed for the study but limiting the literature section to the most recent and relevant papers.

1.2. Motivation and Novelty

From the literature reviewed for the study the most important gap identified is that most of the papers as already mentioned include NLP, CBR, NN and SVM as the ML techniques. There exist few literatures that take the help of other ML algorithms. However, there exists very little research that hybridizes any two ML techniques for analysing the sentiment of the tweets. In order to bridge the gap in the literature, a hybrid NLP and long short term memory (LSTM) based ML method is proposed. The NLP method extracts the keyword from the tweets and the LSTM technique with its feedback connection correlates the tweets to its sentiment.

The remainder of the paper is organized in the following way. Section 2 of the paper discusses the case study and the assumptions considered for solving the problem. Section 3 describes the preliminary concept of the methodology to be used in the paper. Section 4 of the paper discusses the results obtained after applying the proposed methodology in the case study. This section also validates the proposed model. Finally, the paper is concluded in section 5.

2. The case study

A brief description of the case study is discussed in this section of the paper.

2.1. Problem description

Sentiment analysis is the way of getting deeper insights into the opinions and emotions expressed by the social media messages. Political sentiment analysis is the opinions of the crowd about their political leaders and their actions. In the present day, internet users can freely express their opinions on social media and by integrating the views of the users, analysis of the crowd sentiment could be measured. The sentiment analysis is used by political scientists to detect the early crisis and deal with them swiftly. Sentiment analysis also helps for more conscious decisions throughout the political campaign.

Political leaders use social media to build their brand and that is the reason why social media sentiment analysis has become even more valuable in the present political world. The crowd's mood can change at any point of the campaign which can be detected using the help of sentiment analysis. Hence it can help to get a real time insight

of the crowd's mood. The second benefit of sentiment analysis is the speedy escalation system. It helps in early detection of any political crisis.

In the present political scenario it is hard to succeed in politics without fully understanding the electors. Sentiment analysis can be a secret weapon which can target the right demographics by monitoring the overall tone of the conversation. Sentiment analysis is also helpful to monitor the overall trend of a particular political candidate. Also sentiment analysis has the potential to be a powerful weapon in understanding and identifying the emotions of the voters towards their political candidate and what are the topics that can influence their decision.

2.2. The dataset

The dataset used in this study comprises 1.6 million political tweets which have been extracted from Twitter using the Twitter API. All the tweets are annotated with positive, negative and neutral sentiment. The dataset comprises the sentiments, IDs, date on which the tweet is made, the user handle who made the tweet and the text of the tweet. The dataset is extracted from reference number [22].

2.3. Data preprocessing

The process of data preprocessing is a data mining technique that is used to convert the raw data into useful and efficient format. Data preprocessing involves data cleaning, data transformation and data reduction. The data processing in NLP involves tokenization, word stop removal, stemming, segregation and padding.

3. Preliminaries and proposed methodology

In this section of the paper, the preliminary concepts and the proposed methodology used for computing the sentiment analysis of the political trade data set is discussed briefly.

3.1. Natural language processing

In the field of linguistic and artificial intelligence, NLP is used to process and analyse large amounts of natural language data. The goal of NLP is to understand the content of the text and to extract meaningful keywords which can be used to summarize the content [23]. NLP is mostly used in speech recognition, natural language understanding and natural language generation [24]. The problem of sentiment analysis from political tweets is a case of NLP. It helps in extracting the keywords from the text of the tweets.

3.2. Long short term memory (LSTM)

Long short term memory or LSTM is the most cited NN of the 20th century. It is a deep learning model with feedback connections that is capable of processing an entire sequence of data such as speech or video [25]. A typical LSTM unit comprises a cell, an input gate, an output gate and a forget gate. The LSTM networks are most suited to classify processes and make predictions of the time series data. The LSTM models are developed to deal with the vanishing gradient problem which are mostly encountered while dealing with the time series data using the traditional recurrent neural networks [26]. It is mostly applied in speech recognition [27], video games [28] handwriting recognition [29] etc. The reason behind using the LSTM training model is due to the fact that some

words stored in the cloud are predominantly featured in both positive and negative tweets. This could be a problem if other ML models like the Naive Bayes, Support Vector Machine etc. are used.

3.3. Proposed methodology

This paper introduces the hybrid model developed to analyse the sentiment from the political tweets. The NLP part of the model extracts the keywords from the text of the tweets which are then classified using the LSTM method. The steps involved for the proposed methodology are as follows:

Step 1: Tokenization

Tokenization is the process where the tweets are splitted into different tokens or phrases or symbols and words for processing using the NLP method. For example the sentence "I am a good student" is splitted as such "I", "am", "a", "good", "student".

Step 2: Stop words removal

It is the second step for data preprocessing. In this step, words such as I, and, the, for etc. are identified as stop words which are removed from the tweets using NLTK stop word list.

Step 3: Stemming

Stemming is the third step of data preprocessing. In this step, the words are reduced to their base forms. For example: the words such as 'computation', 'computing', 'computes', 'computed' are reduced to the word 'compute'.

Step 4: Segregations

In this step, all the special characters such as [], !, @, #, $, %, &, *, () etc. are removed from the text of the tweets.

Step 5: Padding

In this step, the tweets are truncated to a fixed length. The parent process reduces wastage of memory and fixed length of the tweets is determined by sequence length.

Step 6: Word embedding

In language models, words are represented in such a way that it intends more meaning and for learning the pattern and deriving contextual summary from it. Word embedding is a feature vector representation of words which are used for NLP applications. It is the most popular representation of document vocabulary. It is capable of capturing the context of the words in a document and deriving syntactic similarity or relation with other words. In the proposed model, a pre-trained word embedding namely GloVe Embedding from Stanford AI is used. The pre-trained GloVe embedding gives better insights for a word which can be used for classification. The advantage of GloVe embedding is that unlike other word embedding models it does not rely on local statistics but incorporates Global statistics to obtain word vectors [30].

Step 7: Convolution

In this step, the twitter dataset is convolved into smaller feature vectors.

Step 8: Input for the LSTM unit

Take input the current input, the previous hidden state, and the previous internal cell state.

Step 9: Calculation of the values for the Forget, input, input modulation and output gates.

The values of four different gates LSTM unit is computed in the following way

- For each gate, calculated the parameterized vectors for the current input and the previous hidden state by element-wise multiplication with the concerned vector with the respective weights for each gate.
- Applied the respective activation function for each gate element-wise on the parameterized vectors.

Step 10: Calculation of current internal cell state

Calculated the current internal cell state by first calculating the element-wise multiplication vector of the input gate and the input modulation gate, then calculate the element-wise multiplication vector of the forget gate and the previous internal cell state and then adding the two vectors.

Step 11: Calculation of current hidden state

Calculated the current hidden state by first taking the element-wise hyperbolic tangent of the current internal cell state vector and then performing element-wise multiplication with the output gate.

The flowchart for the proposed methodology is shown in figure (1)

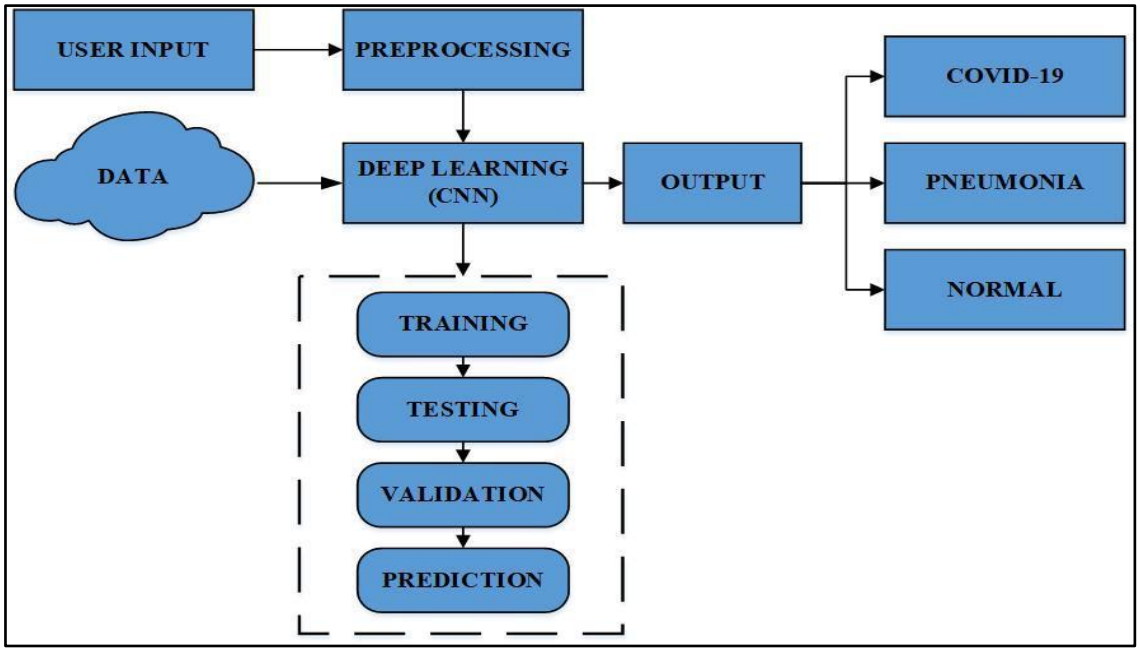

Figure (1): flowchart of the proposed methodology

4. Result and discussion

In this structure of the paper the result obtained from applying the proposed methodology to the case study is discussed briefly.

4.1. Result obtained

The sentiment associated with the tweets in the political Twitter dataset are positive and negative. In order to understand the sentiment distribution of the tweets, the frequency of the sentiment of the tweets are shown in a bar graph in figure (2).

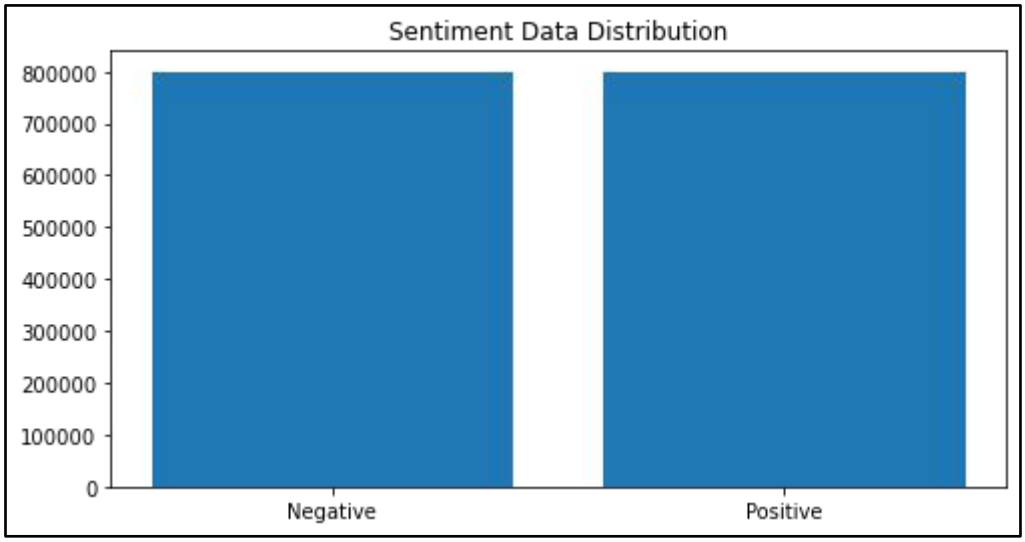

Figure (2): showing the sentiment data distribution

In figure (2), it is observed that there are equal numbers of positive sentiment as well as negative sentiment tweets. Since the number of positive and negative sentiment which are equal hence, we can proceed with the particular data set for the sentiment analysis.

In the study the methodology is coded in Python 3.8.5 and ran on a 64-bit windows 10 system with 8GB RAM and i5, 1.6GHz processor. The different python libraries required to carry out the analysis are listed in table 1.

Table 1: Different libraries imported

Sl. no.	**Library imported**	**Usefulness**	**Sl. no.**	**Library imported**	**Usefulness**
1	Numpy	Used for array operations	**2**	Pandas	Used for data analysis
3	matplotlib	Used for creating static, animated, and interactive visualizations	**4**	nltk	Library used for statistical scrutiny for processing of natural language written in English.
5	Tensorflow	Used for ML and AI	**6**	sklearn	Used for ML

The Twitter dataset is divided in the ratio of 80:20 where 80% of the data is used for training the proposed model and the remaining 20% is used for validation. The proposed model is executed for 10 iterations and the performance of the model is computed by plotting the model accuracy and the model loss. The accuracy curve of the performance of the proposed model is shown in figure (3), on the other hand the loss curve of the performance of the proposed model is shown in figure (4).

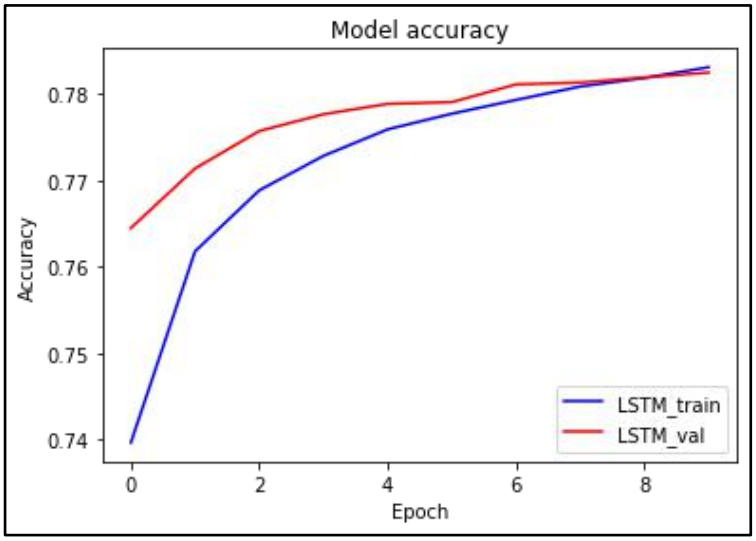

Figure (3): Graph showing the accuracy of the model

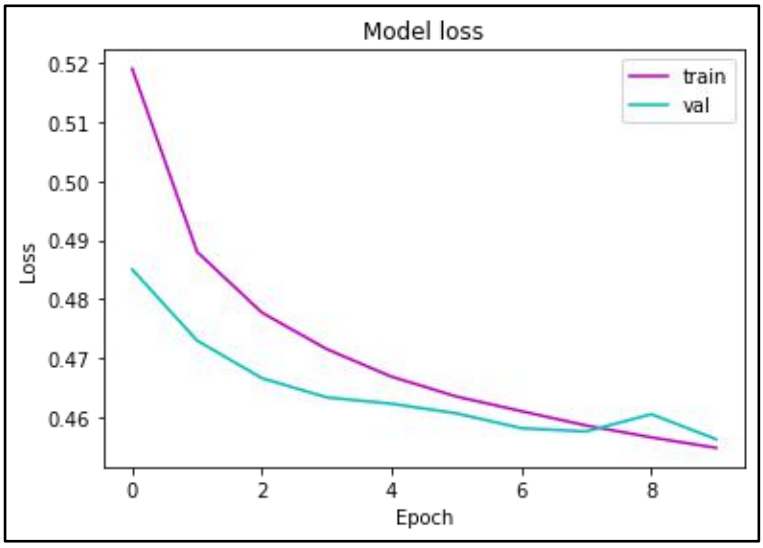

Figure (4): graph showing the the loss of the model

From figure (3) it is observed that the proposed model developed in this paper showed a training and validation accuracy of 74% and 76.4 % respectively in the first epoch. The training and the validation accuracy is improved to 78.4% and 78% respectively in the final epoch. From figure (4), it is observed that the training and validation loss of the proposed model was 0.52 and 0.488 respectively for the first epoch which was subsequently decreased to 0.456 and 0.459 respectively for the final epoch.

A Classification report and a confusion matrix is used to measure the quality of predictions from a classification algorithm. The classification report of the proposed model generated is shown in table 2 and the confusion matrix is shown in figure (5).

Table 2: Classification report of the model

	Precision	**Recall**	**f1-score**	**Support**
Negative	0.79	0.77	0.78	160542
Positive	0.77	0.80	0.78	159458
Accuracy			0.78	320000
Macro average	0.78	0.78	0.78	320000
Weighted average	0.78	0.78	0.78	320000

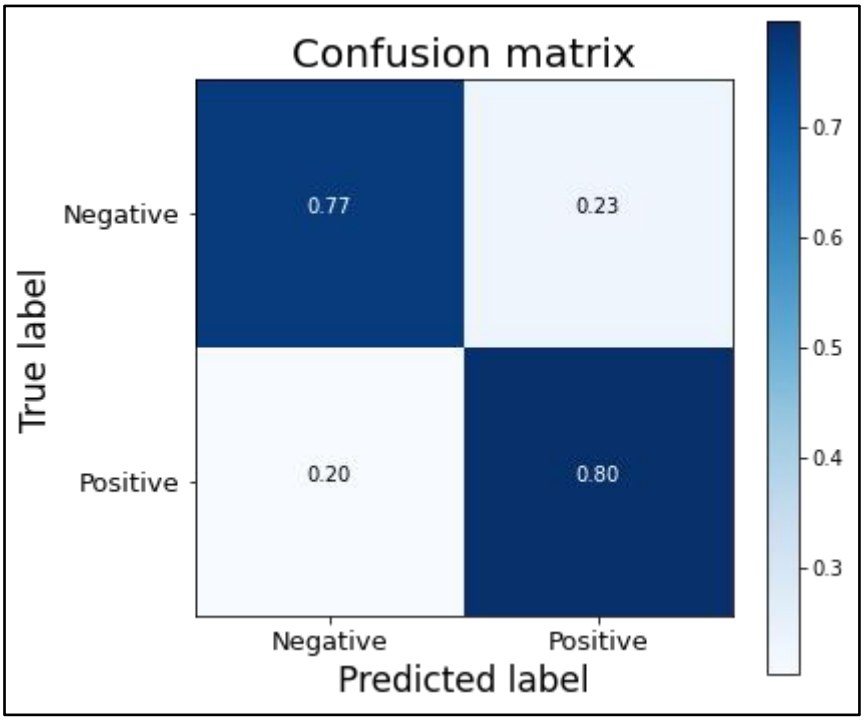

Figure (5): Confusion matrix of the model

From the classification report it is observed that out of the 20% validation dataset, 160542 number of tweets are associated with negative sentiments whereas 159458 tweets are associated with positive sentiment. From the confusion matrix of the proposed model it is observed that 80% of the positive sentiment validation dataset are rightly classified as positive whereas 77% of the negative sentiment validation dataset are rightly classified as negative. On the other hand the proposed model showed 23% of Type 1 error, that is 23% of the negative sentiment validation dataset are wrongly classified as positive. Also the model showed 20% of type 2 error, that is 20% of the positive sentiment validation data set are wrongly classified as negative. From the classification report and

confusion matrix by observing the precision, F1 score and the accuracy of the model which is more than 75%, it can be concluded that the proposed model can be used for sentiment analysis of political tweets.

5. Conclusion

The comprehensive intention of the present study is to propose a hybrid NLP and LSTM based classification model to analyse the sentiments of the political tweets. The political scientist evaluates the sentiment analysis of the common people in order to detect the early crisis and to deal with them swiftly. On the other hand the political strategist evaluates the sentiment analysis to strategize the political campaign of the politicians.

The study presents a hybrid NLP and LSTM based classification model. The NLP method is used for processing the text of the tweets and correlating them with the positive and negative sentiments of the common people. On the other hand the LSTM technique is used to classify the tweets as a positive and negative based on the correlation made by the NLP method. The datasets used for study comprises 1.6 million tweets. The raw datasets are preprocessed by splitting the words in the sentence, reducing the words to their base form, removing special characters and truncating the tweets to a fixed length. Then the word embedding is done to extract more features from the tweets. Word embedding is done using the pretrained GloVe Embedding from Stanford AI. In the next phase convolution operation is done in order to reduce the number of features of the tweets. The processed dataset is used for training the LSTM model. The preprocessed dataset is randomly divided into two parts in the ratio of 80:20 where the 80% portion of the data set is used for training the model and the remaining 20% of the data is used for the validation of the LSTM model.

The hybrid model is coded in Python and ran on a Windows PC. The developed model showed training and validation accuracy of 78.4% and 78% respectively. The developed model also showed training and validation loss of 0.456 and 0.459 respectively . The validation dataset comprises 160542 and 159458 numbers of negative and positive tweets respectively. When the validation dataset is passed through the developed model 80% of the positive and 77% of the negative sentiment tweets are rightly classified as positive and negative respectively. The developed model also showed 23% of Type-I error and 20% of Type-II error. The developed model showed precision of 79% and the computed F1 score of 0.78. Hence it can be concluded that the proposed model can be used to evaluate the sentiments from the political tweets.

Acknowledgement

I would like to express my heartfelt gratitude to all the tutors and mentors of On My Own Technology pvt. ltd. for extending their help in carrying out the particular project. It is because of their help that I am able to conduct the research. I shall remain ever grateful for their help and generosity.

Conflict of interest

The authors would like to declare that the paper is a result of the project work conducted by high school students of Mumbai, Maharashtra, India. The authors would like to declare that no fundings in any form is received for carrying out this research work.

References:

1. He, Y., Saif, H., Wei, Z., & Wong, K. F. (2012). Quantising opinions for political tweets analysis.
2. Pruthi, P., Yadav, A., Abbasi, F., & Toshniwal, D. (2015, June). How has twitter changed the event discussion scenario? a spatio-temporal diffusion analysis. In 2015 IEEE International Congress on Big Data (pp. 733-736). IEEE.
3. Berglez, P. (2016). Few-to-many communication: Public figures' self-promotion on Twitter through joint performances' in small networked constellations. Annales. Series Historia et Sociologia, 26(1), 171-184.
4. Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010, May). Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 4, No. 1, pp. 178-185).
5. Rodríguez-Ibáñez, M., Gimeno-Blanes, F. J., Cuenca-Jiménez, P. M., Soguero-Ruiz, C., & Rojo-Álvarez, J. L. (2021). Sentiment Analysis of Political Tweets from the 2019 Spanish Elections. IEEE Access, 9, 101847-101862.
6. Seçkin, T., & Kilimci, Z. H. (2020, October). The evaluation of 5G technology from sentiment analysis perspective in Twitter. In 2020 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6). IEEE.
7. Maulud, D. H., Ameen, S. Y., Omar, N., Kak, S. F., Rashid, Z. N., Yasin, H. M., ... & Ahmed, D. M. (2021). Review on natural language processing based on different techniques. Asian Journal of Research in Computer Science, 1-17.
8. Jalal, M., Mays, K. K., Guo, L., & Betke, M. (2020). Performance comparison of crowdworkers and nlp tools on named-entity recognition and sentiment analysis of political tweets. arXiv preprint arXiv:2002.04181.
9. Pinto, A., Gonçalo Oliveira, H., & Oliveira Alves, A. (2016). Comparing the performance of different NLP toolkits in formal and social media text. In 5th Symposium on Languages, Applications and Technologies (SLATE'16). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
10. Bird, S. G., & Loper, E. (2004). NLTK: the natural language toolkit. Association for Computational Linguistics.
11. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).
12. Ritter, A., Clark, S., & Etzioni, O. (2011, July). Named entity recognition in tweets: an experimental study. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 1524-1534).
13. Strohmeier, S., & Piazza, F. (2015). Artificial intelligence techniques in human resource management—a conceptual exploration. In Intelligent techniques in engineering management (pp. 149-172). Springer, Cham.
14. Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. Journal of Intelligent Information Systems, 55(1), 51-66.

15. Socher, R., Lin, C. C. Y., Ng, A. Y., & Manning, C. D. (2011, January). Parsing natural scenes and natural language with recursive neural networks. In ICML.
16. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).
17. Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. International Journal of Advanced Computer Science and Applications, 8(6).
18. Ahmad, M., Aftab, S., Bashir, M. S., & Hameed, N. (2018). Sentiment analysis using SVM: a systematic literature review. International Journal of Advanced Computer Science and Applications, 9(2).
19. Taj, S., Shaikh, B. B., & Meghji, A. F. (2019, January). Sentiment analysis of news articles: a lexicon based approach. In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-5). IEEE.
20. Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. Decision Support Systems, 53(4), 680-688.
21. Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.
22. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.
23. Chowdhary, K. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.
24. Allen, J. F. (2003). Natural language processing. In Encyclopedia of computer science (pp. 1218-1222).
25. Hochreiter, S. JA1 4 rgen Schmidhuber (1997).“Long Short-Term Memory”. Neural Computation, 9(8).
26. Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen [in German] Diploma thesis. TU Münich.
27. Zia, T., & Zahid, U. (2019). Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. International Journal of Speech Technology, 22(1), 21-30.
28. Stanford, S. DeepMind's AI, AlphaStar Showcases Significant Progress Towards AGI. Medium ML Memoirs, 2019. Alphastar has a" deep LSTM core.
29. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2008). A novel connectionist system for unconstrained handwriting recognition. IEEE transactions on pattern analysis and machine intelligence, 31(5), 855-868.
30. Eke, C. I., Norman, A., Shuib, L., Fatokun, F. B., & Omame, I. (2020, March). The significance of global vectors representation in sarcasm analysis. In 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS) (pp. 1-7). IEEE.