IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10556

NLP Based Covid-19 Sentiment Classification and fake news detection using ML

Thayaba Nausheen A[1], Sujatha B R[2]

M.Tech DECS Student, Department of Electronics and Communication Engineering, Malnad College of Engineering, Hassan, India[1]

Professor, Department of Electronics and Communication Engineering, Malnad College of Engineering, Hassan, India[2]

Abstract: In this age, the Internet has empowered the flow of thoughts and data and has thus expanded the knowledge base among individuals. Be that as it may, this has its own disadvantages in terms of false and fake data. The need to uncover such bogus data and disdain discourse during this COVID-19 pandemic has never been more important. Data mining is fundamentally used to identify relevant required data available on the internet or any data source. Consolidating the information mining using different methods like content mining, NLP and computational insight, we can arrange Corona virus tweets as great, awful or unbiased. The objective of this work is on the characterization of feelings of 'Corona virus tweets' information assembled from Twitter. Moreover, we can characterize the Covid-19 Texts as phony or not. To improve characterization we are utilizing AI methods for improving the effectiveness and quality of the proposed approach**.**

Keywords: NLP, Naïve Bayes, Covid-19 data, Fake news detection, Sentiment Classification

I. INTRODUCTION

The existing COVID-19 pandemic has, straight forwardly or by implication, influenced each life on this planet. The primary human instances of COVID-19, likewise named as SARS-CoV-2, were at first found in Wuhan city, China, in December 2019. In any case, as the infection was not contained, the cases spread quickly from individual to individual inside neighbourhoods throughout the Planet. As of March 11 2020, the WHO chief general pronounced COVID-19 a pandemic. It was not simply a health related emergency any longer, as it can dubiously affect each area and each person. The Covid pandemic has required an exceptional change in the working of government personnel and medical services experts. Fig. 1 shows the number of tweets posted related to pandemic [1],

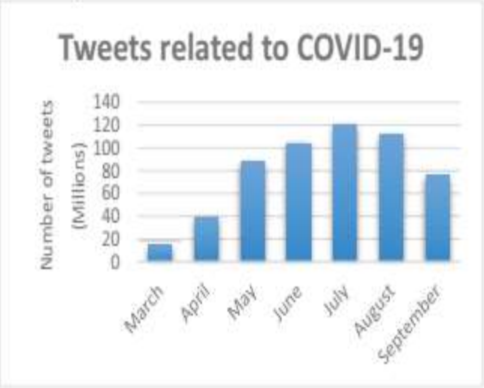


Fig. 1 Representation of the tweets posted during the period of March 20, 2020, to September 21, 2020

The consistently expanding amount of 'Bogus News' has immersed the reality that the individuals were supposed to know to take precautionary. There has been a call for handling this flood of bogus data in the current pandemic so that the correct real data is available for the general public.

There have been various articles and innovative advances in utilizing distinctive gear of Natural Language Processing (NLP) to identify and help battle the problems of transmission of phony news, reports, and surprisingly malevolent and disdain remarks. Additionally, NLP can also be utilized to identify the recorded events occurring in COVID-19 data. NLP can utilize text acknowledgment to decide designs in counterfeit matters, which is used to sort out those issues utilization comparative study.

This work is licensed under a Creative Commons Attribution 4.0 International License

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10556

The objective of this work is to find the fake news which is spread in various media and classify them as fake or not. Section Ⅱ specifies related work; Section Ⅲ specifies principles of Natural Language Processing; Section Ⅳ shows system design; Section Ⅴ indicates software implementation followed by results; lastly, section Ⅵ presents conclusion of our work.

II. RELATED WORK

Several works related to NLP and data mining is found in literature. S. A. Cammel *et al* [1] proposed a method for calculating sentiments and frequency values for development of progress. This model can be used for capturing similar data from various hospitals. In this model unsupervised learning is used which is used in representation of patients feedback hence avoiding human bias.

M. U. Salur and I. Aydin [2] proposed a deep learning model such as M-Hybrid, CNN etc. for representation of data in text format especially which is written in Turkish, Arabic which are difficult to understand. These models have accuracy up to 82.14%. This model also uses learning models of embedding and classifies the text with respect to sentiments.

Statistical investigation, conclusion examination, etymological signs examination are been used for detecting the true news (Double-dealing speculations) in the work proposed by C Zhang [3]. The Model proposed by W. Haitao is adaptable and permits clients to module data sets considered dependable by them, creating valuable highlights for short content. This methodology utilizes CNN's alongside non-straight sliding and N-gram examination to fundamentally improve short content order [4].

Creators distinguish the accompanying systems .Noisy words eliminated for better execution [5][6]. Model highlights the methodology for catching the design of other languages apart from English. Creators recognize the accompanying philosophies such as Natural language handling (NLP) of biomedical content and identification of vague words [7].

Deep learning models are used to conquer issues presented by NLP. Model uses a live mining stage to consolidate helper highlights from relating news stories for better execution [8]. Tweets gathered from March 13 to March 21, 2020, are pre-processed by using NLP and by using this method tweets which has negative meaning can be recognized [9]. Some information was haphazardly examined, physically named, and algorithms such as Random forest have been applied. Various analyses on chosen highlights for each sort of data was done [10].

M. Cinelli et al proposed a work in which various data related to COVID-19 has been extracted with the assistance of NLP [11]. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III.PRINCIPLES OF NLP

Natural language processing (NLP) is subset of artificial intelligence (AI). It helps in interaction between man and machine. NLP has different applications such as utilized for grammar recognizable proof, notion investigation. It is likewise used to decipher languages, Chat bots, additionally utilized in observing various online media.

IV. SYSTEM DESIGN

The System design is as shown in fig. 2

Stage`1: This stage takes Twitter survey information for performing a feeling examination. After taking survey information it performs supposition sentence extraction for creating assessment targets.

Stage 2: This stage removes a portion of discourse labelling as a contribution from stage 1 and distinguishes estimation express. This stage additionally registers the conclusion score and produces a highlight vector. Highlight vector is shaped dependent on a sentence (survey).

Stage 3: This stage performs two activities, for example, supposition extremity order which gives after effect of conclusions of positive, negative or unbiased.

Result translation gives results as either phony or not.

This work is licensed under a Creative Commons Attribution 4.0 International License

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10556

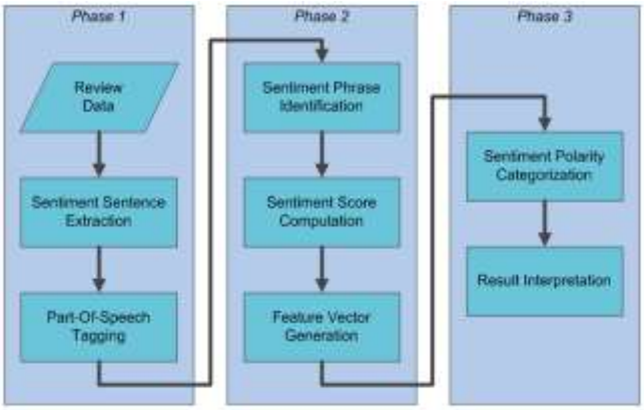

Phase 1
Phase 2
Phase 3
Review Data
Sentiment Sentence Extraction
Part-Of-Speech Tagging
Sentiment Phrase Identification
Sentiment Score Computation
Feature Vector Generation
Sentiment Polarity Categorization
Result Interpretation

Fig. 2 System design architecture

V. IMPLEMENTATION

A. Sentiment sentences extraction and POS tagging

Sentiment sentences extraction and POS labelling the target substance ought to be considered for assessment examination [3]. Rather than eliminating target content, in our investigation, all emotional substance was separated for future examination. The passionate substance contains all inclination sentences that contain words which gives negative meaning. POS taggers have been made to orchestrate words subject to their syntactic highlights.

B. Negation phrases identification

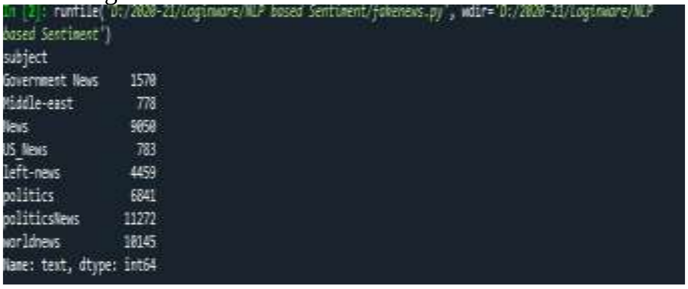
Words, for example, modifiers and action words can pass on inverse estimation with assistance of negative prefixes. For example, consider the sentence: For example, consider the sentence: "Getting vaccinated may not decrease Covid cases. Here “decrease Covid cases” is negative phrase.

C. Naïve Bayes Algorithm

Naive Bayes Algorithm relies upon the Bayesian theory. It is particularly applicable when the dimensionality of the information sources is high. Limit evaluation for Naive Bayes models uses the procedure for most likelihood cases.

VI. RESULTS

Subject wise data count and their graphical representation of the news article which are represented in Fig. 4 and Fig. 5 respectively. Fig. 6 represents the most commonly used words in the article which is combination of both fake and true news. Fig7 represents the cloud of words which comprises of only fake news whereas Fig. 8 represents the cloud which consists of only true news and Fig. 9 indicates Naïve Bayes Classification Report which represents the accuracy of true and fake news obtained by using this algorithm. Fig. 10, Fig. 11 and Fig. 12 represents the various use cases of sentiment analysis such as negative, positive and neutral respectively. Fig. 13 represents the comparison of accuracy by using algorithms such as Naïve Bayes and Linear regression which indicates that Naive Bayes algorithm is more efficient when compared to Linear Regression.



```
In [2]: runfile('D:/2020-21/Loginware/NLP based Sentiment/fakenews.py', wdir='D:/2020-21/Loginware/NLP
based Sentiment')
subject
Government News    1570
Middle-east         778
News               9050
US_News             783
left-news          4459
politics           6841
politicsNews      11272
worldnews         10145
Name: text, dtype: int64
```

Fig. 4 Subject-wise Data Count

This work is licensed under a Creative Commons Attribution 4.0 International License

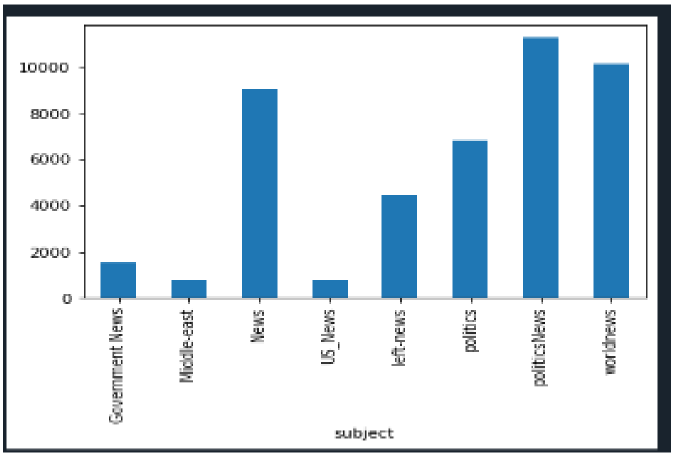


Fig. 5 Graphical View of Subject-Wise Count

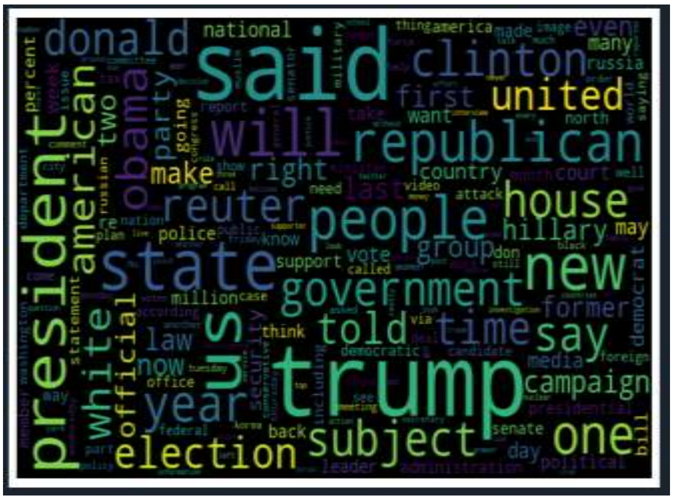


Fig. 6 Word cloud News visualization

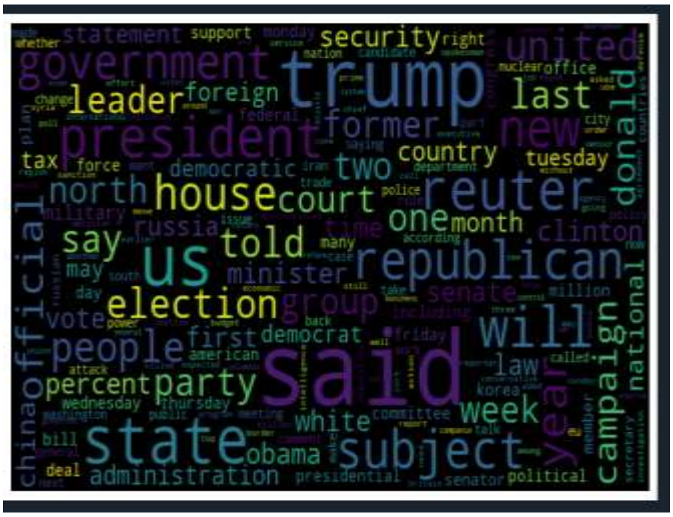


Fig. 7 Word cloud of Fake News

This work is licensed under a Creative Commons Attribution 4.0 International License

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10556

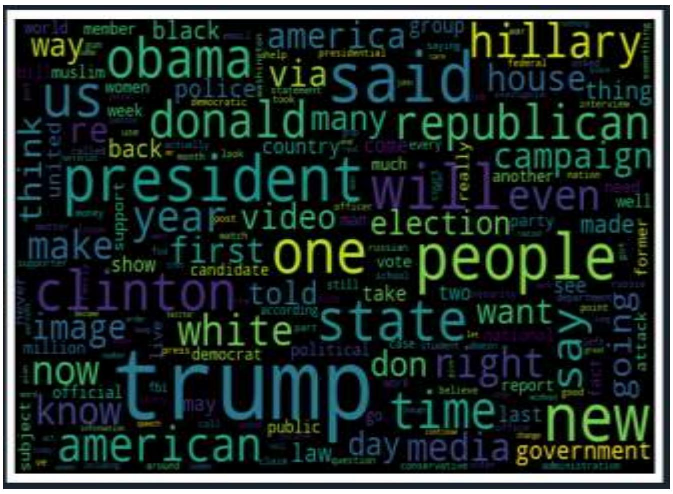
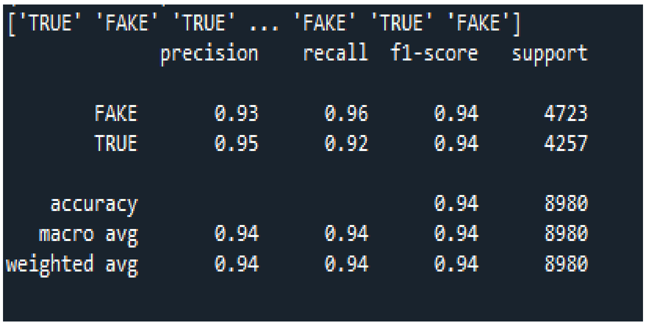


Fig. 8 Word cloud of True News

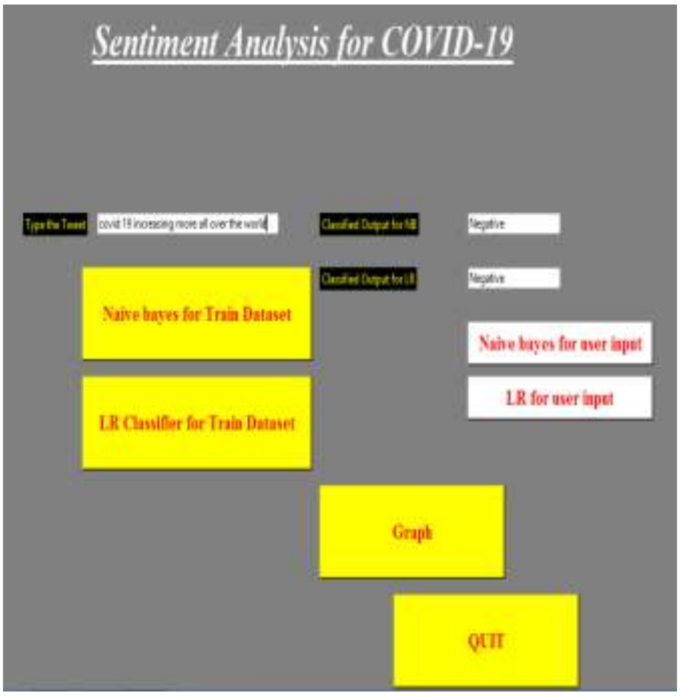


```
['TRUE' 'FAKE' 'TRUE' ... 'FAKE' 'TRUE' 'FAKE']
              precision    recall  f1-score   support

        FAKE       0.93      0.96      0.94      4723
        TRUE       0.95      0.92      0.94      4257

    accuracy                           0.94      8980
   macro avg       0.94      0.94      0.94      8980
weighted avg       0.94      0.94      0.94      8980
```

Fig. 9 Naïve Bayes Classification Report



Sentiment Analysis for COVID-19

Type the Tweet: covid 19 increasing more all over the world

Classified Output for NB: Negative

Classified Output for LR: Negative

Naive bayes for Train Dataset
LR Classifier for Train Dataset
Naive bayes for user input
LR for user input
Graph
QUIT

Fig. 10 Sentiment analysis for twitter input- Case Negative

This work is licensed under a Creative Commons Attribution 4.0 International License

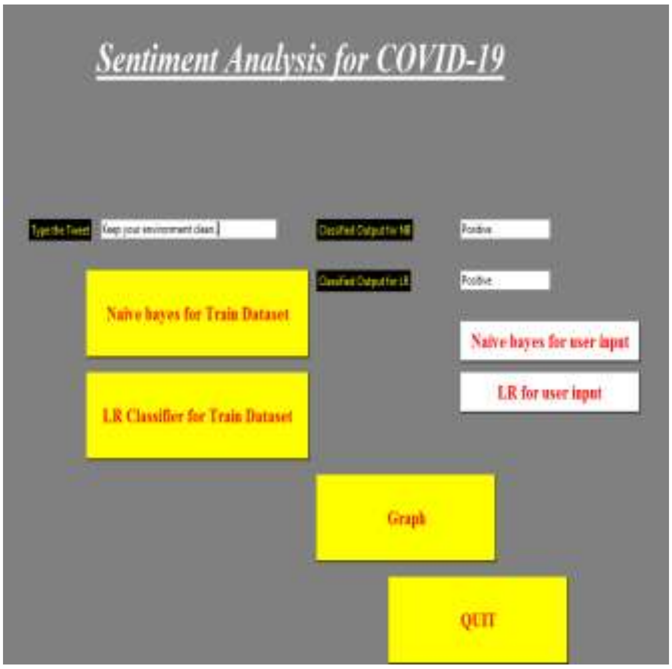
IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10556

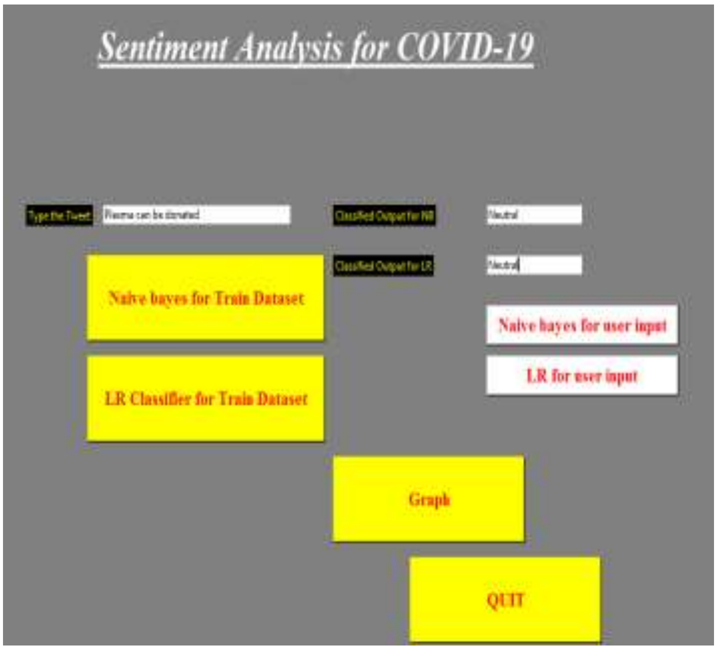

Sentiment Analysis for COVID-19
Naive bayes for Train Dataset
LR Classifier for Train Dataset
Naive bayes for user input
LR for user input
Graph
QUIT

Type the Tweet: Keep your environment clean.

Classified Output for NB: Positive

Classified Output for LR: Positive

Fig. 11 Sentiment analysis for twitter input- Case Positive


Sentiment Analysis for COVID-19
Naive bayes for Train Dataset
LR Classifier for Train Dataset
Naive bayes for user input
LR for user input
Graph
QUIT

Type the Tweet: Plasma can be donated

Classified Output for NB: Neutral

Classified Output for LR: Neutral

Fig. 12 Sentiment analysis for twitter input- Case Neutral

This work is licensed under a Creative Commons Attribution 4.0 International License

DOI 10.17148/IJARCCE.2021.10556

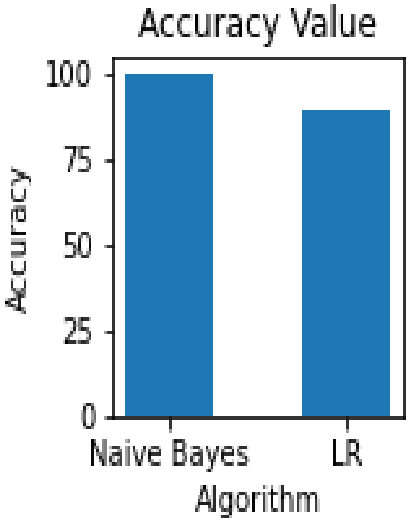


Fig. 13 Comparision of algorithms

VII. CONCLUSION

The spread of phony news and events happening during Covid pandemic are analysed using Natural Language Processing methods. A set of strategies to acquire customary critical thinking approaches about specific issues by experimenting on dataset is discussed. The goal of this work is to manage individual analysts to carry out and enhance the current best in class models utilizing NLP, to assist them in applying these ideas in COVID pandemic. The location of phony news and disdain remarks that occurred in the tweets of COVID-19 has to be refreshed continuously in different datasets.

REFERENCES

[1] S. A. Cammel *et al.*, "How to automatically turn patient experience free-text responses into actionable insights: A natural language programming (NLP) approach," *BMC Med. Inform. Decis. Mak*, vol. 20, no. 1, pp. 1–10, 2020, DOI: 10.1186/s12911-020-1104-5.

[2] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," *IEEE Access*, vol.8, pp.58080–58093, 2020, DOI:10.1109/ACCESS.2020.2982538.

[3] C. Zhang, A. Gupta, C. Kauten, A. V. Deokar, and X. Qin, "Detecting fake news for reducing misinformation risks using analytics approaches, "*Eur. J. Oper. Res.*, vol. 279, no.3, pp.1036–1052, 2019, DOI: 10.1016/j.ejor.2019.06.022.

[4] W. Haitao, H. Jie, Z. Xiaohong, and L. Shufen, "A short text classification method based on n-gram and CNN," *Chinese J. Electron.*, vol. 29, no. 2, pp. 248–254, 2020, DOI: 10.1049/cje.2020.01.001.

[5] V. Agarwal, H. P. Sultana, S. Malhotra, and A.Sarkar, "Analysis of Classifiers for Fake News Detection," *Procedia Comput. Sci.*, vol. 165, no. 2019, pp. 377–383, 2019, DOI: 10.1016/j.procs.2020.01.035.

[6] A. Alhelbawy, M. Lattimer, U. Kruschwitz, C. Fox, and M. Poesio, "An NLP-Powered Human Rights Monitoring Platform," *Expert Syst. Appl.*, vol. 153, 2020, DOI: 10.1016/j.eswa.2020.113365.

[7] Z. Li, F. Yang, and Y. Luo, "Context Embedding Based on Bi-LSTM in Semi-Supervised Biomedical Word Sense Disambiguation," *IEEE Access*, vol. 7, pp. 72928–72935, 2019, DOI:10.1109/ACCESS.2019.2912584.

[8] S. Deepak and B. Chitturi, "Deep neural approach to Fake-News identification," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 2236–2243, 2020, DOI:10.1016/j.procs.2020.03.276

[9] S. A. Chun, A. C. Y. Li, A. Toliyat, and J. Geller, "Tracking citizen's concerns during COVID-19 pandemic," *ACM Int. Conf. Proceeding Ser.*, pp. 322–323, 2020, DOI: 10.1145/3396956.3397000.

[10] L. Li *et al.*, "Characterizing the Propagation of Situational Information in Social Media during COVID-19 Epidemic: A Case Study on Weibo," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 2, pp. 556–562, 2020, DOI: 10.1109/TCSS.2020.2980007.

[11] M. Cinelli *et al.*, "The COVID-19 Social Media Infodemic," pp. 1–18, 2020.

[12] H. Jelodar, Y. Wang, R. Orji, and H. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Corona virus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," *IEEE J. Biomed. Heal .Informatics*, vol. 2194, no. c, pp. 1–1, 2020, DOI:10.1109/jbhi.2020.3001216.

[13] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y.Zhang, and S. Zannettou, "'Go eat a bat, Chang!': An Early Look on the Emergence of Sinophobic Behaviour on Web Communities in the Face ofCOVID-19," vol. 2, 2020.

[14] R. Pandey *et al.*, "A Machine Learning Application for Raising WASH Awareness in the Times of COVID-19 Pandemic," 2020.

[15] E. De Santis, A. Martino, and A. Rizzi, "An Infoveillance System for Detecting and Tracking Relevant Topics from Italian Tweets During the COVID-19 Event," *IEEE Access*, vol. 8, pp. 132527–132538, 2020, DOI: 10.1109/access.2020.3010033.

This work is licensed under a Creative Commons Attribution 4.0 International License