INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Hybrid Approach to Analyse the Public Sentiment on Covid-19 Tweets

Radha Krishna Jana[1]*, **Dharmpal Singh**[1], **Saikat Maity**[2], **Hrithik Paul**[1]

1 Department of Computer Science & Engineering, JIS University, Kolkata, India
2 Department of Computer Science & Engineering, Sister Nivedita University, Kolkata, India

OPEN ACCESS

Received: 26-11-2023

Accepted: 13-01-2024

Published: 08-02-2024

Citation: Jana RK, Singh D, Maity S, Paul H (2024) A Hybrid Approach to Analyse the Public Sentiment on Covid-19 Tweets . Indian Journal of Science and Technology 17(7): 610-616. https://doi.org/10.17485/IJST/v17i7.3017

* **Corresponding author**.

radhakrishnajana@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The objective of this study is to introduce a hybrid model for analyzing the people sentiment on covid-19 tweets. **Methods:** We used a total no. of 27,500 datasets, 70% of the data sets for training and reserved the other 30% for testing. Due to this separation 19,250 samples are used for training, the remaining 8,250 were used to evaluate the accuracy of the test. This paper proposes a technique for sentiment analysis that integrates deep learning, genetic algorithms (GA), and social media sentiment. For more accuracy and performance, we here suggested a hybrid genetic algorithm-based model. A hybrid model is created by assembling the LSTM model and providing it to the genetic algorithm architecture. **Findings:** LSTM with a genetic model better than LSTM without genetic model. The accuracy of our suggested model is 96.40%. **Novelty:** The accuracy of the LSTM model for sentiment analysis is 91%. The accuracy of the proposed model is 96.40%. The proposed model is more accurate for sentiment prediction.

Keywords: Social network perception; Crossover; Mutation; LSTM; NLP; GA

1 Introduction

We are busy with social networking sites in present day. We are posting our comments, emotions and opinions on different events, product etc. On social networking sites, different types of people are form groups. They share opinions in this group. Expressing one's thoughts on social media can be a tricky affair, as it can be difficult to discern whether the sentiment is positive or negative. Social media is a platform where individuals share their life experiences through tweets, making it challenging to understand each person's unique circumstances. This is where sentiment analysis comes into play, as it analyzes the tweet and provides a brief classification as either positive or negative. However, enhancing the performance of sentiment analysis is a complex task, as an incorrect classification can lead to inaccurate results. As a result, a new approach has been developed to improve the performance of the sentiment analysis model for this particular dataset. Homophily groups are formed using same minded people[(1)]. All the sentiment and behaviour are needed for analysis. There are various types of methods for sentiment analysis. Hybrid method gives the better accuracy than other methods[(2)]. Now a days people are busy in social networking site. They commented on different types of events. They give their opinion on different products and different decisions.

This types opinion, comment provides the social network perception for the particular topic. Social network perception plays an important role on marketing or statistical analysis[3]. Social networking sites has important role in many areas. There are various tools for learning purposes. This work discussed student perceptions about various education tools in social networking[4].

Genetic algorithm (GA) is an optimization algorithm. There are some encoding steps. Different working steps in genetic algorithms are initialization, reproduction, crossover operation, inversion, and mutation. Different types of operators in genetic algorithms are crossover, mutation, and selection[5].

Here we discuss different genetic algorithm approaches on different problem. This study works for sentiment detection. This study proposed a hybrid model with genetic and Neural Network. This work gives better accuracy than the old model[6]. A challenging task for gold investors is gold price prediction. Most research work is pursued using conventional techniques. They depend on economic indicators. This study proposed a text mining approach[7].

This algorithm is used for sentiment analysis of big data. This classifier gives better performance than the current classifier. The proposed algorithm applied on an online large dataset[8]. Feature selection of online sentiment is a challenging task. There are several algorithms for feature selection. This work gives better accuracy than other algorithms[9]. This study was proposed on a social network of Arabic. The proposed model analyzing the customer behaviour. The main task of this study is feature analysis[10]. This algorithm proposed for adaptive lexicon generation. Text classification is the main work of this study[11]. This work performed on analysis of public transport. The proposed model work on Indonesian text[12]. This work proposed a Social Media Pandemic Sentiment Model. They used LSTM and Genetic algorithm for detecting the sentiment during pandemic[13].

There is a challenge for identifying the polarity of textual data. There are several approaches proposed. This study proposed a good model for this work[14]. During covid-19 world of people send their views in social media. This time different initiatives taken by administrators of different nations of the world. This study showed the sentiment of people during this time[15]. This study developed a model to analyze the public sentiment on covid-19. After analyzing the public sentiment, the administrator can guide the people to fight the pandemic during this period[16]. This worked to help to analyze the public sentiment on covid-19. The proposed model showed better accuracy than other models[17]. People post their comments on social media in different format. A hybrid approach constitutes for sentiment analysis during covid-19[18]. This study builds a prediction model Covid-19 cases. They are used five algorithms. Random forest gives the better accuracy than other algorithm[19].

All the recent studies do not properly analyze the user sentiment. There are several algorithms and models used. But shows less accuracy. The main contribution of our research work as follows:

- We analyze the user sentiment using their text in social media.
- We proposed a hybrid model using LSTM and Genetic Algorithm.
- We used covid-19 tweets. We applied LSTM without genetic algorithm and LSTM with genetic algorithm on the same datasets. Our proposed hybrid model (LSTM with GA) showed the highest accuracy of all recent works.

Our research related methodology is described in section 2. Experiments and discussion are provided on section 3. Section 4 presented conclusion and future study.

2 Methodology

For this study, we used open datasets obtained from Kaggle. Specifically, our testing focused on Twitter data sets, which are purchased randomly. These twitter datasets are provided in CSV format and included a total no. of 27,500 data.

Before the next step, the first step required is data preprocessing where the first thing is removing null values from the entire data set. After this subtraction, the class values are converted to their corresponding integer values. Text features are transformed into tokens and these tokens are converted into sequence of integers as the pad sequence. After this careful data preprocessing step, we divided the datasets into training and testing.

To run our experiments, we used 70% of the data sets for training and reserved the other 30% for testing. Due to this separation 19,250 samples are used for training, the remaining 8,250 were used to evaluate the accuracy of the test. The decision to use the CSV format was based on its timeliness, making it especially convenient for creating, writing, and acquiring information.

The genetic algorithm then evaluates the fitness of each individual and selects those with the highest scores for reproduction. The selected individuals undergo a crossover process to create new offspring and the genetic mutation process to introduce randomness and diversity in the population. The new population is evaluated, and the highest-scoring individuals are selected for the next iteration. This process continues until an optimal solution is found.

Genetic algorithm consists of various parts such as Initialization, Fitness Function, Selection, Crossover, Mutation, and Termination. Initialization is the first step and involves randomly creating a population of individuals. Fitness Function or Fitness Assignment evaluates the fitness of each individual on the basis of their genes. Selection phase is where the best performing individuals are selected for reproduction. Crossover is a process of combining two individuals to generate a new offspring with better genes.

Mutation is an essential part of the genetic algorithm, and it occurs after the fitness assignment stage. It is the process of randomly changing the values of a chromosome or a gene in a population. This helps to increase the accuracy and reduce the loss of the model. It can involve changing the data type, removing, or adding a feature to the model, or changing the value of a parameter.

Once a mutation has been performed, the process continues until the stopping criteria has been met. At this point, the model has been trained with updated parameters and optimized features, resulting in a much higher accuracy and lower loss. Once the stopping criteria is met, the genetic algorithm stops executing and the model is trained with the new parameters, forming a hybrid model.

After forming the hybrid model then we use a test dataset to predict sentiment values. Below, Figure 1 shows the whole architecture of our proposed work.

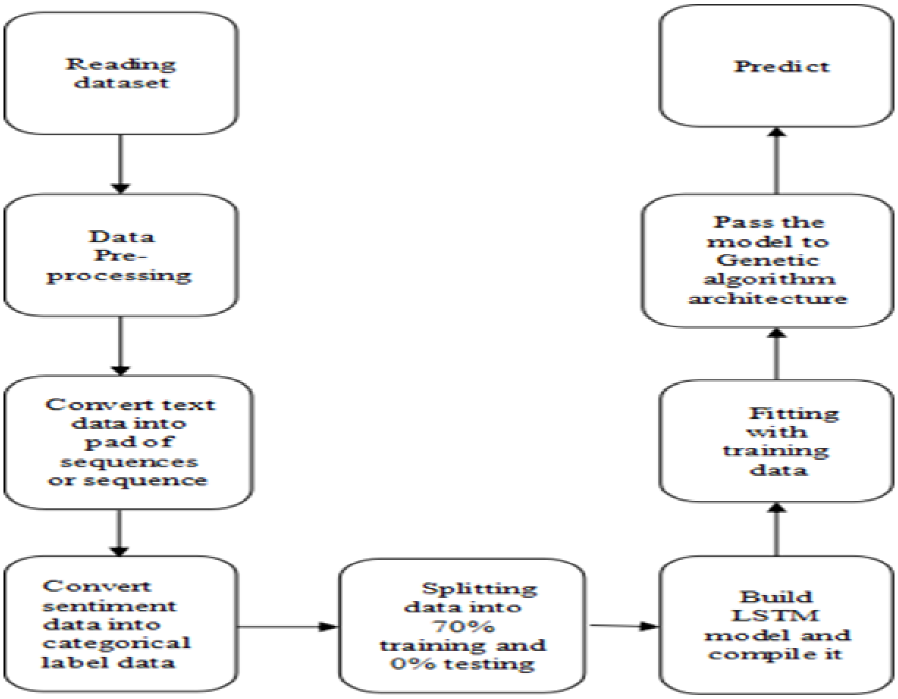
Reading dataset
Data Pre-processing
Convert text data into pad of sequences or sequence
Convert sentiment data into categorical label data
Splitting data into 70% training and 0% testing
Build LSTM model and compile it
Fitting with training data
Pass the model to Genetic algorithm architecture
Predict

Fig 1. Proposed work

The given image shows that the process begins with reading the data and preprocessing it, which involves cleaning the data. Afterwards, the words are counted and tokenized. The data is then converted into sequences of integers, while sentiment data is transformed into labels. Afterwards, the data is split into 70% training and 30% testing data. Finally, a Long Short-Term Memory (LSTM) model is built and passed to a genetic algorithm architecture, forming a hybrid model.

3 Results and Discussion

In this study, we employed freely available open-source datasets from Kaggle. For our experimentation, we utilized Twitter datasets, which were obtained at no cost. The Twitter datasets were provided in CSV format and consisted of a total of 27,500 data entries. To conduct our experiments, we divided the datasets into 70% for training and 30% for testing, resulting in 19,250 instances used for training and 8,250 instances used for testing accuracy evaluation. The decision to use the CSV format was based on its simplicity, making it convenient for creating, writing, and accessing content, particularly with Python's built-in functions.

Cleaning and preprocessing data is an essential step for any deep learning task, as it helps to ensure accurate and reliable results from the model. In this case, the data was first cleaned by removing special characters, digits, and unnecessary symbols, as well as converting words to their root form for easier interpretation. Tokenizing the data into individual words, and then counting them, allowed for further analysis of the data. The data is then subjected to sentiment analysis, which transforms sentiment terms like ”positive,” ”negative,” into categorical data like [1,0] or [0,1] which is subsequently used as the target data. After that, text data is transformed into an integer sequence. After that, the category and numerical data are divided into training and testing sets, with the remaining data being utilized 70% for training and the remaining 30% for testing. The data splitting is depicted in the graphic below.

The top num-words-1 most frequent words are taken into account, and only words recognized by the tokenizer are used after the data has been converted into a series of numbers. The length of the tokenized data was utilized to determine the vocab size. The pad_Sequences method is then used to convert the list of sequences into a 2D Numpy array of shape (numsamples, numtimesteps). This made it possible to create deep learning models that were more reliable and accurate. We'll then start creating models after that. The first layer is composed of the primary LSTM layer, the output layer, the spatial dropout layer, and the embedding layer. The table below shows the parameters for each layer:

Table 1. Different Parameter in each layer

Layers, dropout	Input dim	Output dim	Activation	Dropout Rate	units
Embedding	vocab_size	Vector length, 32	-	-	-
Spatial Dropout	-	-	-	0.25	-
LSTM	-	-	-	0.5	50
Dropout	-	-	-	0.2	-
Output layer	-	-	-	sigmoid	1

It is clear from the following table that the range values for various levels vary. Additionally, a dropout layer has been employed to solve the overfitting problem. Spatial-Dropout 0.25 has been proposed for the 50-unit primary LSTM layer of neurons. Vocabulary_size and vector_size are embedded in the input layer's embedding layer, and 1 unit has been utilised with sigmoid activation as the output layer since it predicts a single value. Then we will compile our model using “Adam” optimizer. We have used other optimizers also like “adagrad”,”rmsprop”, “adadelte” etc. But among all of them, we have got the best accuracy with the lowest loss in “Adam”.

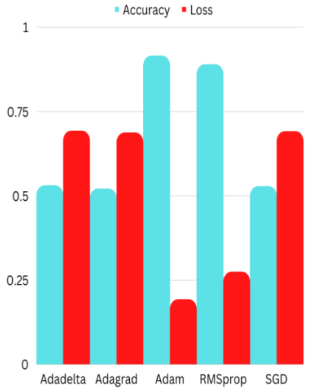

Fig 2. Optimizers comparison

The x-axis in the image above (Figure 2) represents all experimental optimisations, while the y-axis represents accuracy value. The blue bar for accuracy and the red bar for loss are shown in this bar graph. Among the optimizers we used were ”Adam,” ”adadelta,” ”adagrad,” ”rmsprop,” and ”sgd.” It is clear that among all of these optimizers, ”Adam” has the highest accuracy and the lowest loss. The accuracy and loss levels are shown, respectively, by the hues blue and red.

We start by building the model using the best optimizers, and after that, we train the model using a particular epoch. Here, the graph below displays the accuracy level broken down into epochs:

As seen in the preceding picture, accuracy falls between 91 and 90, and it stays the same even when the period size is raised by up to 20. Therefore, we will use a genetic algorithm in neural networks to improve accuracy and make the model more reliable

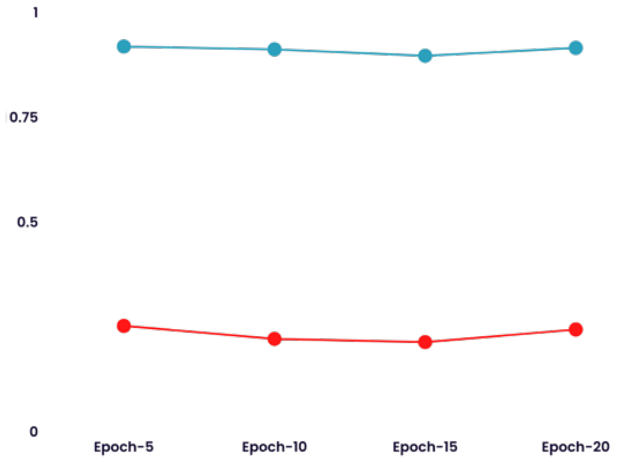

Fig 3. Epoch comparison

in the event of an accurate prediction. Before genetic algorithm using, the accuracy is steady. Which is not increasing after a certain value. Even if we increase the iteration, the accuracy still remains in the certain range like 90 to 91. So, without genetic algorithms the model accuracy likely remains the same. But after using the genetic algorithm with deep neural networks, the model accuracy increases. Even so, it becomes so good in training.

The LSTM model achieved an accuracy of 91% and 22% loss before the incorporation of Genetic Algorithm. However, after integrating GA in the model. The hybrid approach resulted in a significant improvement, achieving an accuracy of 96% and 9% loss.

Table 2. Optimizers comparison of proposed model

Model	Accuracy	Loss
adam	0.96	0.123
rmsprop	0.953	0.23
adadelta	0.53	0.54
adagrad	0.76	0.45
sgd	0.56	0.5444

Table 3. Neural networks weights and parameters

Model	Embedding layer	Dropout	Dense	LSTM	Spatial dropout
LSTM (Genetic algorithm)	vcab_size=15317units and iput_length=200	0.2	1 units	50 units	0.25

As shown in Table 4, we have compared the results of recent work. The proposed work has 70% accuracy[19]. The authors used the BERT model and got 82.90% accuracy[20]. The proposed experiment showed 96.40% accuracy[21]. The accuracy of the GA-DNN model is 85%[22]. AEGA model has an accuracy of 78.60%[23]. Laura Imanuela Mustamu et al.[24] got an accuracy of 69.52%. Our model has the highest accuracy of 96.40%.

We choose to assess the model's sentiment categorization performance using actual test data, in keeping with the objectives stated in our study. Both of the models used in our experiment carried different results. In the preceding table, all of the texts—even the ones with incorrect labels—are consistently classified as positive when using LSTM without Genetic Algorithm (GA). All four classes, however, are correct once LSTM and GA are combined. When texts reflect negative sentiment, the model predicts the negative class properly; when texts reflect good emotion, it predicts the ”positive” class correctly.

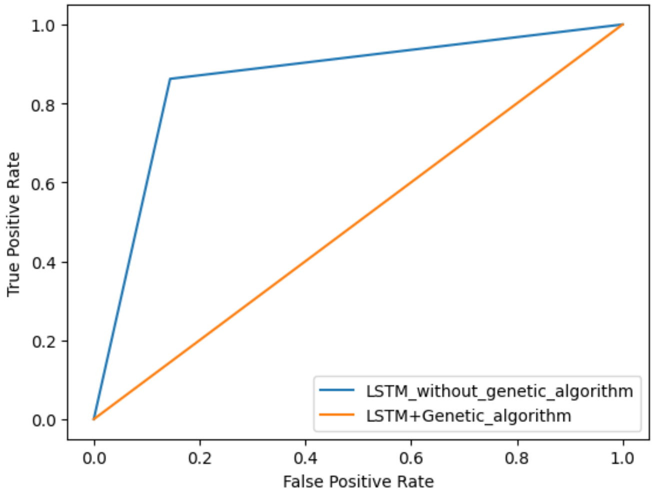

Fig 4. LSTM & Genetic Algorithm

Table 4. Comparison of Recent work

Recent Work	Models Used	Accuracy (%)
Kanak Mahor et al. [25]	Transfer-based BERT	70%
Yuchen Chai et al. [20]	BERT	82.90%
S Ashika Parvin et al. [21]	HCNN-BiLSTM	96%
Omar Al-Harbi et al. [22]	GA-DNN	85%
Gyananjaya Tripathy et al. [23]	AEGA	78.60%
Laura Imanuela Mustamu et al. [24]	RBF and GA-SVM	69.52%
Our Proposed Model	LSTM-With GA	96.40%

Table 5. (a, b) Comparison of LSTM-Without GA between LSTM With GA of Positive, Negative

(a) LSTM-Without GA	
Text	**Sentiment**
today launch woman aid annual impact report special covid19 report report provides insights	Positive
break florida state student test posit covid19 florida state locate tallahasse Florida	Positive
help fan request help arrange bed posit father view tweet	Positive
heart break stat u one wealthiest country world happens covid19 e	Positive
(b) LSTM With GA	
today launch woman aid annual impact report special covid19 report report provides insights	Positive
break florida state student test posit covid19 Florida state local tallahasse florida	Negative
help fan request help arrange bed posit father view tweet	Positive
heart break stat u one wealthiest country world happens covid19 e	Negative

4 Conclusion

The relevant accuracy of the LSTM model for sentiment analysis without GA is 91%, which falls short of the more accurate sentiment prediction. A better model is essential for better prediction analysis.

From that vantage point, the hybrid LSTM model used in our proposed genetic method provides greater accuracy. A hybrid model is created by building a Long Short-Term Memory model and adding genetic algorithm architecture. The accuracy of our suggested model is 96.40%, which is significantly higher than before. And in the next few days, the hybrid model will gain more significance in the field of study. Future efforts will focus more on hybrid approaches to improve sentiment analysis's performance. We come to the conclusion that this will serve as a model for better performance than other models in the future.

Despite the improved performance of the model, certain limitations still exist in certain scenarios. Specifically, when the text is incomplete, such as ”I am,” ”done,” ”forsake,” etc., the model fails to predict the correct class. This implies that the model performs well only when the text is fully completed with an accurate meaning, even taking into account spelling errors. However, if the text is too short or incomplete, the model sometimes fails to classify the correct output.

References

1) Jana RK, Maity S, Maiti S. An Empirical Study of Sentiment and Behavioural Analysis using Homophily Effect in Social Network. In: and others, editor. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE. 2022;p. 1508–1515. Available from: https://doi.org/10.1109/ICICCS53718.2022.9788407.
2) Jana RK, Maity S. An Accuracy Based Comparative Study on Different Techniques and Challenges for Sentiment Analysis. In: Pervasive Computing and Social Networking;vol. 475 of Lecture Notes in Networks and Systems. Singapore. Springer. 2023;p. 601–619. Available from: https://doi.org/10.1007/978-981-19-2840-6_46.
3) Ahmed C, Elkorany A, Elsayed E. Prediction of customer's perception in social networks by integrating sentiment analysis and machine learning. *Journal of Intelligent Information Systems*. 2023;60(3):829–851. Available from: https://doi.org/10.1007/s10844-022-00756-y.
4) Sanwal T, Yadav S, Avasthi S, Prakash A, Tyagi M. Social Media and Networking Applications in the Education Sector. In: 2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON). IEEE. 2023;p. 1–6. Available from: https://doi.org/10.1109/DELCON57910.2023.10127547.
5) Katoch S, Chauhan SS, Kumar V. A review on genetic algorithm: past, present, and future. *Multimedia tools and applications*. 2021;80:8091–8126. Available from: https://doi.org/10.1007/s11042-020-10139-6.
6) Sangule S, Phulre S. Sentiment Detection Using Genetic Feature Vector And Neural Network Model. *International Journal of Advanced Research in Engineering and Technology (IJARET)*. 2020;11(12):2726–2734. Available from: https://iaeme.com/MasterAdmin/Journal_uploads/IJARET/VOLUME_11_ISSUE_12/IJARET_11_12_257.pdf.
7) Yuan FC, Lee CH, Chiu C. Using Market Sentiment Analysis and Genetic Algorithm-Based Least Squares Support Vector Regression to Predict Gold Prices. *International Journal of Computational Intelligence Systems*. 2020;13(1):234–246. Available from: https://doi.org/10.2991/ijcis.d.200214.002.
8) Merlin DJA, Kumar DV. Perceptive Genetic Algorithm-Based Wolf Inspired Classifier For Big Sentiment Data Analysis. *Journal Of Theoretical And Applied Information Technology*. 2022;100(16):5021–5031. Available from: https://www.jatit.org/volumes/Vol100No16/13Vol100No16.pdf.
9) Rafdi A, Mawengkang H, Efendi S. Sentiment Analysis Using Naive Bayes Algorithm with Feature Selection Particle Swarm Optimization (PSO) and Genetic Algorithm. *International Journal of Advances in Data and Information Systems*. 2021;2(2):96–104. Available from: https://doi.org/10.25008/ijadis.v2i2.1224.
10) Al-Qudah DA, Al-Zoubi AM, Castillo-Valdivieso PA, Faris H. Sentiment Analysis for e-Payment Service Providers Using Evolutionary eXtreme Gradient Boosting. *IEEE Access* . 2020;8:189930–189944. Available from: https://doi.org/10.1109/ACCESS.2020.3032216.
11) Al-Shabi MA. Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *International Journal of Computer Science and Network Security*. 2020;20(1):51–57. Available from: http://paper.ijcsns.org/07_book/202001/20200107.pdf.
12) Aryanti R, Saryoko A, Junaidi A, Marlina S, Wahyudin, Nurmalia L. Comparing Classification Algorithm With Genetic Algorithm In Public Transport Analysis. In: International Conference on Advanced Information Scientific Development (ICAISD) ;vol. 1641 of Journal of Physics: Conference Series. IOP Publishing. 2020;p. 1–6. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1641/1/012017.
13) Rani P, Shokeen J, Majithia A, Agarwal A, Bhatghare A, Malhotra J. Designing an LSTM and Genetic Algorithm-based Sentiment Analysis Model for COVID-19. In: Proceedings of Data Analytics and Management;vol. 91 of Lecture Notes on Data Engineering and Communications Technologies. Singapore. Springer. 2022;p. 209–216. Available from: https://doi.org/10.1007/978-981-16-6285-0_17.
14) Sravya G, Sreedevi M. Genetic Optimization in Hybrid Level Sentiment Analysis for Opinion Classification. *International Journal of Advanced Trends in Computer Science and Engineering*. 2020;9(2):1440–1445. Available from: https://www.warse.org/IJATCSE/static/pdf/file/ijatcse81922020.pdf.
15) Wang J, Fan Y, Palacios J, Chai Y, Guetta-Jeanrenaud N, Obradovich N, et al. Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nature Human Behaviour*. 2022;6(3):349–358. Available from: https://doi.org/10.1038/s41562-022-01312-y.
16) Luu TJP, Follmann R. The relationship between sentiment score and COVID-19 cases in the United States. *Journal of Information Science*. 2023;49(6):1615–1630. Available from: https://doi.org/10.1177/01655515211068167.
17) Arbane M, Benlamri R, Brik Y, Alahmar AD. Social media-based COVID-19 sentiment classification model using Bi-LSTM. *Expert Systems with Applications*. 2023;212:1–9. Available from: https://doi.org/10.1016/j.eswa.2022.118710.
18) Bashar MK. A Hybrid Approach to Explore Public Sentiments on COVID-19. *SN Computer Science*. 2022;3(3):1–19. Available from: https://doi.org/10.1007/s42979-022-01112-1.
19) Gupta VK, Gupta AK, Kumar D, Sardana A. Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Mining and Analytics*. 2021;4(2):116–123. Available from: https://doi.org/10.26599/BDMA.2020.9020016.
20) Chai Y, Kakkar D, Palacios J, Zheng S. Twitter Sentiment Geographical Index Dataset. *Scientific Data*. 2023;10(1):1–12. Available from: https://doi.org/10.1038/s41597-023-02572-7.
21) Parvin SA, Sumathi M, Barani R. A Novel Approach to Classify Sentiments on Different Datasets Using Hybrid Approaches of Sentiment Analysis. *Indian Journal of Science and Technology*. 2023;16(44):3962–3970. Available from: https://doi.org/10.17485/IJST/v16i44.2498.
22) Al-Harbi O, Hamed A, Alzoubi M. A Deep Neural Network Optimized by a Genetic Algorithm to Improve Arabic Sentiment Classification. *Ingénierie des systèmes d information*. 2023;28(1):67–75. Available from: https://doi.org/10.18280/isi.280107.
23) Tripathy G, Sharaff A. AEGA: enhanced feature selection based on ANOVA and extended genetic algorithm for online customer review analysis. *The Journal of Supercomputing*. 2023;79(12):13180–13209. Available from: https://doi.org/10.1007/s11227-023-05179-2.
24) Mustamu LI, Sibaroni Y. Fuel Increase Sentiment Analysis Using Support Vector Machine With Particle Swarm Optimization And Genetic Algorithm As Feature Selection. *Jurnal Teknik Informatika (Jutif)*. 2023;4(3):521–528. Available from: https://doi.org/10.52436/1.jutif.2023.4.3.881.
25) Mahor K, Manjhvar AK. Public Sentiment Assessment of Coronavirus-Specific Tweets using a Transformer-based BERT Classifier. In: 2022 International Conference on Edge Computing and Applications (ICECAA). IEEE. 2022. Available from: https://doi.org/10.1109/ICECAA55415.2022.9936448.