An Efficient CSPK-FCM Explainable Artificial Intelligence Model on COVID-19 Data to Predict the Emotion Using Topic Modeling

Priya C. and Durai Raj Vincent P. M. *

School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India
Email: priya.2017@vitstudent.ac.in (P.C.)
*Correspondence: pmvincent@vit.ac.in (D.R.V.P.M.)

Abstract—**Incessant COVID-19 pandemic negatively impacts nations throughout the globe. It is necessary to determine how people react to public health interventions and understand their concerns. Twitter is a social media platform that has emerged as a tool for disseminating information, debating concepts, and reviewing or commenting on global issues. This study applies Explainable Artificial Intelligence (XAI) methods, like Cosine Similarity and Polynomial Kernel-centered Fuzzy C-Means (CSPK-FCM) centered topic modeling and Fuzzy Logic with Improved Long Short-Term Memory (FL-ILSTM) centered Sentiment Analysis to COVID-19 data on Twitter. The proposed model has five major steps: preprocessing, feature extraction, term weighting, topic modeling (clustering), and classification. Twitter comments relating to the COVID-19 pandemic are initially collected from publicly accessible websites. The collected data are then preprocessed to remove irrelevant information, namely, noises. The Feature Extraction phase is then performed by extracting emoticon and non-emoticon features. The extracted feature dataset is scored: the Term Frequency Inverse Document Frequency-Chi-Square (TFIDF-CHI) method is utilized for non-emoticon, and the score for the emoticon is assigned based on a few criteria. For Topic modeling, the TFIDF-CHI scores are provided to the CSPK-FCM clustering algorithm, which groups the most frequently discussed topics throughout COVID-19. FL-ILSTM executes the Sentiment analysis of clustered topics and emoticon features. It has extraordinary performance when compared to other methodologies.**

Keywords—**COVID-19, topic modeling, sentiment analysis, twitter sentiment analysis, Explainable Artificial Intelligence (XAI), fuzzy logic, Long Short-Term Memory (LSTM)**

I. Introduction

COVID-19 (SARS-CoV-2) [1] was initially identified in China's Hubei Province around December 2019. The COVID-19 virus [2], which has rapidly spread globally, causes a severe respiratory disease. The COVID-19 outbreak was declared a pandemic by the World Health Organization on March 11, 2020 [3, 4]. Not only is COVID-19 a worldwide pandemic, but it also resulted in an "infodemic (surplus information)" and a "disinfodemic (the disinformation whirling amongst the COVID-19 pandemic)" [5]. Due to the swift advancement of the scientific foundation for treating the disease and rapid decline in employment within affected economies, the government and PH agencies have tried to mobilize ushered in a period of enormous and multifaceted social change, a consequence of these ongoing shifts that evolve rapidly and are often significantly, albeit inaccurately, depicted by millions of voices on social media [6]. Social media is a famous channel for news and information in the well-timed media environment, with one-third of the global population using Social Media (SM) and two-thirds utilizing the internet [7]. Numerous studies demonstrated that SM plays a crucial role as a data source for identifying outbreaks and for understanding public attitudes and behaviors during a crisis to support crisis communication and health promotion messaging [8–10].

However, the difficulty of the analysis stems from the SM data source. Social media messages are extremely noisy and idiosyncratic, and the volume of incoming data is too large to analyze manually. Therefore, automatic methods are required to extract meaningful insights [11]. Topic Modeling (TM) and Sentiment Analysis (SA) are the most widely utilized methods in PH for detecting issues and people's opinions. In addition, they are utilized for understanding COVID-19-associated problems [12]. Topic models can facilitate search, navigation, and knowledge discovery in massive document compilations [13, 14]. The information in the set will be represented utilizing an unsupervised Machine Learning (ML) method. By sorting associated terms together in topics and detecting SM patterns, the TM's effect helps enhance classification [15]. The opinions of individuals via social networks, namely Facebook, Twitter, and Instagram, are comprehended by SA [16, 17]. Twitter is one of the most well-liked SM platforms, with an average of 330 million active users (monthly) sharing content on the platform in 2019. Twitter users can publish publicly accessible posts (called tweets), making Twitter a rich data source for monitoring social phenomena and PH concerns [18]. The proposed TM and SA models for

Manuscript received June 19, 2023; revised August 12, 2023; accepted August 21, 2023; published December 14, 2023.

doi: 10.12720/jait.14.6.1390-1402

COVID-19 are evaluated in Section II. However, restricted insight exists into the major topics discussed and the common public's sentiment eventually. Second, the number of works that fall under both TM and SA is minimal, as most works have concentrated on TM or SA. During the pandemic, it is essential to examine both topics and sentiments discussed by the people. Third, several works examine emotions using ML classifier, which has the following drawbacks: (1) interpretation of results, (2) requires adequate time to permit the algorithms to study and develop sufficiently to fulfill their purpose with a considerable quantity of accuracy and relevance, and (3) they are highly vulnerable to errors.

This study proposes a CSPK-FCM clustering model for conducting a TM on Twitter data that finds topics and eventually tracks the topic change. In addition, an XAI-centered Sentiment Classification (SC) which is similar to FL-ILSTM is proposed for evaluating the user's sentiments and manifestations (comments, hashtags, posts, tweets) of the Twitter SM platform based on the major trends in Natural Language Processing. The two research objectives are elucidating online concerning COVID-19 associated discussion themes and analyzing COVID-19 related sentiments. This study's results can assist researchers in determining what information about the pandemic is pertinent and how people respond to it. Therefore, this information can be utilized by researchers and authorities to identify the pandemic's significant aspects that can guide better action and communication policies toward the population.

This paper is categorized into four sections. Section II offers the associated work. Section III briefly describes the proposed work. Section IV proffers the proposed method's results and discussion. Lastly, Section V presents the conclusion.

II. Literature Review

This section reviews recently proposed methods of TM and SA. The techniques are evaluated based on the utilized methodology, attained results, and limitations.

Abd-Alrazaq *et al.* [19] analyzed the tweets in English to determine the most popular subjects on Twitter related to the COVID-19 epidemic from February 2, 2020, to March 15, 2020. Latent Dirichlet Allocation (LDA) was used for TM. They discovered 12 topics organized into four main topics: the virus's root, its outlet, the virus's impact on the public, nations, and the financial system, and preventative measures. The mean sentiment for 10 topics was positive, with only two topics displaying negative sentiment (deaths resulting from COVID-19 and augmented racism). However, the tweets from private accounts were not collected for analysis. Therefore, the findings cannot represent every COVID-19-related topic discussed by Twitter users.

Mackey *et al.* [20] implemented a Bi-term Topic Modelling for tweets related to COVID-19 on March 3, 2020. This model, divided into topic clusters, included the same word-associated topics regarding symptoms, testing, and recovery. The tweets were sorted into five key categories: illness reporting with a concurrent deficit of testing, recovery discourse, corroboration of a negative COVID-19 result following testing, and users recollecting symptoms and starting to question whether they had previously been infected with COVID-19. The cross-validation of the veracity of user-created comments with other data sources (such as confirmed case reports, extra survey data, death records, data on other diseases with related symptoms, or electronic medical records) was not conducted.

Wang *et al.* [21] employed the unsupervised Bidirectional Encoder Representations from the Transformers (BERT) model for categorizing SA sentiment categories (positive, neutral, and negative). The TF-IDF model was utilized for summarizing the post's topics. Trend and thematic analysis are performed to identify negative sentiment characteristics. People believed that the virus's origin, production activity, symptoms, and PH control were the four aspects of COVID-19 that concern them. The SC model and findings provide constructive instructions for governments globally in making efficient PH protection decisions.

Imran *et al.* [22] examined the citizen's reactions to COVID-19 from disparate cultures and people's sentiments regarding the subsequent actions taken by diverse nations. Pre-trained embedding models were employed using deep learning Long Short-Term Memory (LSTM) architecture that attained top-notch accuracy upon the Sentiment 140 dataset. The emotional tweet dataset was utilized to identify sentiment polarity and emotions of users' tweets on Twitter. The work contrasted the disparate word embeddings' performance namely GloVe, and BERT. The Recurrent Neural Network (RNN)'s different variants containing LSTM, Bi-LSTM, and GRU were utilized for SC. However, the deep neural network's performance, specifically Convolutional Neural Networks (CNN), and its variants were not evaluated. In addition, the employed word embedding did not capture the word context.

Jelodar *et al.* [23] utilized automated extraction of COVID-19 associated discussions from SM and a Natural Language Process (NLP) centered on TM to identify different issues correlated to COVID-19 from public opinions regarding TM and SA. In addition, it was investigated how to utilize LSTM RNN for SC of COVID-19 comments. Based on the results, the model achieved an accuracy of 81.15% and was similar to the numerous famous ML algorithms for COVID-19 SC. The findings can assist in enhancing practical strategies for PH services and interventions related to COVID-19.

Ordun *et al*. [24] demonstrated five different techniques for measuring the individuality of COVID-19 tweets topics, keywords, features, information distribution, and network behaviors. Initially, pattern matching was utilized. Second, TM through LDA was utilized to generate 20 topics related to case spread among Healthcare (HC) workers and Personal Protective Equipment (PPE). Third, the Uniform Manifold Approximation and Projection (UMAP) algorithm were employed to enhance the comprehension of significant themes in the corpus and assess the generated topic's quality that detected distinct clustering behavior of different topics. Fourth, retweets were utilized to estimate the retweeting rate. Fifth, the networks of COVID-19 retweeting communities were planned with retweeting times ranging from rapid to extended. Since the number of nodes reduced, every network's density eventually increased.

Xue *et al.* [25] examined Twitter users' conversations and psychological responses toward COVID-19 using ML techniques. 1.9 million Tweets (written in English) correlated to COVID-19 were gathered from January 23 to March 7, 2020, for analysis. Eleven main topics were detected and categorized into ten themes, including "updates regarding confirmed cases," "COVID-19 associated death," "cases out of China (globally)," "COVID-19 outbreak in South Korea," "outbreak's initial signs in New York," "Diamond Princess cruise," "economic effect," "Preventative measures," "authorities," and "supply chain." As shown by SA, the fear of COVID-19's unfamiliar nature was dominant in every topic. The method for collecting Twitter data sampled only trending '19' hashtags as search words. Eventually, a few other hashtags emerged as trending terms (new) aimed at Twitter users for sorting topics.

Chakriswaran *et al.* [26] investigated Twitter data on movie reviews and provided a comprehensive survey report on Ontology based-sentiment analysis, lexicon-based sentiment analysis, and machine learning models with the experimental results. The used tweets were written in English.

Ren *et al.* [27] worked on a real-time dataset for analysis. They aimed to overcome challenges such as data sparsity and the difficulty in capturing and determining special events. Consequently, they implement self-learning and weight modification to ensure the general and burst sensitivity of the model using a Deep Deterministic Policy Gradient (DDPG) framework with Long Short-Term Memory (LSTM) networks.

Leow *et al.* [28] proposed two models, the Sentimental All-Weather (SAW) and Sentimental MPT (SMPT) models, capturing the most recent market conditions using Twitter sentiments via Google's Bidirectional Transformer (BERT) model. The models were optimized using a genetic algorithm to achieve various goals, such as maximizing cumulative returns and reducing volatility. Their proposed models outperformed the following benchmarks: the buy-and-hold SPY index, the MPT model, and the CRB model for an All-Weather Portfolio regarding common portfolio performance measures, such as the Sharpe ratio, cumulative returns, and value-at-risk.

Haung *et al.* [29] intended to clarify how attitudes changed in various levels of interaction during various phases of blended learning. This study combined text mining and ENA techniques to achieve this. The combined method uncovered six types of feelings and six-dimensional interactions from the qualitative data collected during the discussions in the blended learning activities and the five learning phases identified by LSTM.

A well-known classification technique with a wide range of applications is K-Nearest Neighbours (KNN) [30]. The use of the Euclidean distance as the similarity metric, the randomly chosen neighbourhood size k, the computational difficulty of high-dimensional data, and the use of the simple majority voting rule in class determination limit KNN despite its simplicity, effectiveness, and robustness. So, the author developed the Centroid Displacement-based KNN technique, where centroid displacement is employed for class identification, in an effort to address the final problem. The Ensemble Centroid Displacement-based KNN presented in this study is a straightforward yet effective variation on their earlier work that makes use of the homogeneity of test cases' close neighbors.

A popular supervised machine learning technique in numerous domains is KNN [31]. Despite being straightforward, efficient, and robust, KNN is restricted to using the Euclidean distance as the similarity metric, picking the neighborhood size k at random, dealing with high-dimensional data's computational challenge, and using the simple majority voting rule. We proposed the Centroid Displacement-based KNN (CDNN), a variation of the KNN used in classification, in an effort to address the last problem. The CDNN uses centroid displacement to determine the classes. In this work, author show an implementation of CDNN for scikit-learn, a popular machine learning package for Python, as well as a thorough comparison of CDNN's performance with several KNN variations in scikit-lear.

The machine learning model (SVM) resulted in an accuracy of 90%, whereas the Aspect based-ontology approach produced an accuracy of 83%, and the Lexicon-based approach TFIDF produced an accuracy of 85% when analyzing the sentiment analysis. These results are beneficial for quickly understanding the overall feedback or summary from various kinds of people.

III. Materials and Methods

Social networks are a popular tool for advertising and disseminating ideas and initializing individual opinions. In general, examining social network content can offer insight into society and the world. Understanding people's emotions is imperative in the situation resulting from COVID-19. Utilizing CSPK-FCM and FL-ILSTM, a TM and SA are utilized to process COVID-19 Twitter data (English Tweets) in this paper. The Twitter data during the COVID-19 pandemic over the period from January to August 2020 was initially gathered from the publicly accessible database. Thereafter, the amassed data was preprocessed by conducting URL removal, repeated word removal, and removal of special characters. Using the TF-MICF method, Term Weightage (TW) of the preprocessed data was performed after preprocessing. The data comprises the emoticon features that are also extracted as the preprocessed data. The score value for the extracted emoticon features was allotted based on several criteria. After that, the term weights of the non-emoticon features were taken and provided into CSPK-FCM clustering for TM. Employing the XAI model, SA was conducted after TM. FL-ROLSTM was used to categorize the emotions of the topics that are deemed throughout COVID-19. Fig. 1 exhibits the proposed method's framework.

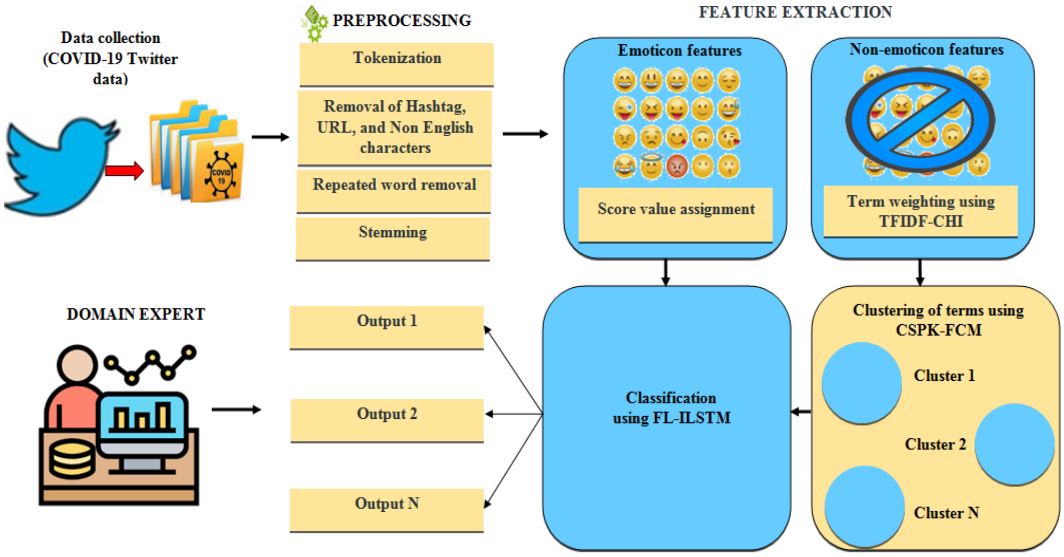


Figure 1. Framework of the proposed method.

A. *Preprocessing*

The raw data was preprocessed to ensure quality. The preprocessing steps are as follows:

- Tokenization was performed to divide a whole sentence into smaller units named tokens.
- Since the symbols (hashtag) or the URLs did not provide the message analysis, the hashtag symbol and its content (for example, #COVID-19), @users, and URLs as of the messages were eliminated.
- Since the study focuses on the message's analysis in English, every non-English character (non-ASCII characters) was eliminated.
- The repeated words in the gathered tweets were eliminated. For instance, "sooooo terrified" was changed to so terrified.
- As for the dataset, the punctuations, special characters, and numbers were removed because they did not assist in identifying the profanity comments.
- The stemming process was conducted in which a word was reduced to its word stem, affixing to suffixes and prefixes.

B. *Feature Extraction*

The FE phase was performed after preprocessing. In this case, emoticon (smileys) and non-emoticon features (texts) are the two features that were extracted. Additionally, a score value was allotted for both features. The emoticon features' score value was allotted based on the following criteria. Initially, the extracted smiley was detected for their classes (happiness, sadness, fear, disgust, surprise, and anger). The score for the extracted smiley was allocated in the gamut (8–10) if the identified smiley was in the happy class. Similarly, the score value was allocated for every smiley class. The score was fixed in (–5 to –6) and (–8 to –10) for sadness and fear. The scores were allotted as (2.5–3.5), (6–8), and (1–2) for disgust, surprise, and anger. Afterward, the score value was allotted based on the TFIDF-CHI model for non-emoticon features that were explained in the section below.

C. *Term Weighting*

TFIDF-CHI is a new weighting scheme utilized for attaining a favorable TW (feature score). The text Feature Selection (FS) process is facilitated and enhanced by effectively differentiating between informative and uninformative text features. The standard weight scheme is TFIDF. Assigning TW over every review by merging the CHI with the TFIDF is typically influenced by the presented model, which concentrates on repairing the existing TFIDF weakness. Normally, a static TW method is Chi-Square (CHI), which attempts to catch the intuition that the best terms t_j for the class c_i are the ones distributed most in a different manner in the sets of disparate classes. The CHI terms can be calculated utilizing Eq. (1).

$$C_{HI}\left(t_j,c_i\right)=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{R_N\left(d\left(t_j.c_i\right).d\left(t''_j.c''_i\right)-d\left(t''_j.c_i\right).d\left(t_j.c''_i\right)\right)^2}{d\left(t_j\right).d\left(t''_j\right).d\left(c_i\right).d\left(c''_i\right)} \quad (1)$$

where R_N is the total reviews required for training, $d(t_j)$ is the total reviews including term t_j, $d(t''_j)$ is the total reviews that are not comprising the term t_j, $d(c_i)$ is the overall reviews which are members of class c_i, $d(c''_i)$ is the number of reviews that are not members of class c_i, $d(t_j.c_i)$ is the number of reviews in class c_i that include the term t_j, $d(t''_j.c''_i)$ is the number of reviews not belonging to class c_i and not encompassing term t_j, $d(t''_j.c_i)$ is the total reviews i.e., the members of class c_i but not containing term t_j, and $d(t_j.c''_i)$ signifies the number of reviews not belonging to class c_i but including term t_j.

CHI is merged with TFIDF to overcome the drawback of TFIDF, which is specified as:

$$W_{TFIDF-CHI}\left(t_j\right)=TF\left(t_j,d_k\right)\left[1+\log\left[\frac{R_N}{d\left(t_j\right)}\right]\right]C_{HI}\left(t_j,c_j\right) \qquad (2)$$

where $TF(t_j,d_k)$ is the t_j occurrence frequency in the document d_k, the 2nd term in Eq. (2) is the t_j IDF (class frequency), and the third term signifies the CHI scheme.

This score is deemed a final score for TW in non-emoticon features.

D. Topic Modeling

The TM for the data features extraction (non-emoticon) was performed after FE by creating the clusters using the CSPK-FCM algorithm. Fuzzy C-Means (FCM) clustering is a soft clustering method in which every data point is assigned a likelihood or probability score related to that cluster. When creating the clusters, the Euclidean Distance (ED) is utilized by the conventional FCM for measuring the distance between the cluster center and data point. The two data vectors might have a lesser distance analogized to the pair of data vectors comprising the same attribute values if they have no common attribute values. The difference is treated evenly, sometimes not fulfilling the necessities. Thus, a Cosine Similarity (CS) based on the distance computation was introduced into the FCM that decides how similar the data objects are irrespective of their size to enhance FCM's performance. The CS is helpful because although the two similar data objects are far apart by the ED due to their size, they can still contain a small angle between them. Besides, the Polynomial Kernel (PK) function in distance computation was utilized through the CS that maps the comparatively low-dimensional data into a high-dimensional feature space, which can recognize non-linear relationships in the features. The integration between CS and PK for FCM is called CSPK-FCM. The FCM is based on minimizing the subsequent Objective Function (OF):

$$F_{ob}=\sum_{i=1}^{N}\sum_{j=1}^{C}p_{ij}^{w}\left\|z_i-c_j\right\|^2,\quad 1\le w<\infty \qquad (3)$$

where w is a real number greater when contrasted to 1, p_{ij} is the membership degree of z_i in the cluster j, z_i is the i th of dimensional estimated data, c_j is the cluster's d-dimension center which denotes any norm expressing the ED betwixt any estimated data and the center.

The similarity between two vectors is assessed rather than calculating ED between two vectors by d coupling PK function into CS measurement, which is defined in Eq. (4):

$$S_{ij}=\cos\left(_{z_i,c_j}\right)=\frac{Z_i,C_j}{\sqrt{Z_i.Z_i}\sqrt{C_j.C_j}}\rightarrow\frac{y''\left(z_i,c_j\right)}{\sqrt{y''\left(z_i,z_i\right)}.\sqrt{y''\left(c_j,c_j\right)}} \qquad (4)$$

Eq. (4) is further simplified as:

$$S_{ij}=\frac{\partial\left(z_i\cdot c_j+\beta\right)^k}{\sqrt{\left(\partial z_i\cdot z_i+\beta\right)^k}\cdot\sqrt{\left(\partial c_j\cdot c_j+\beta\right)^k}} \qquad (5)$$

where ∂ and β are parameters, and k implies the degree.

This S_{ij} value can be utilized instead of $\left\|z_i-c_j\right\|^2$ in Eq. (3). The fuzzy partitioning was performed with the membership p_{ij} using the OF's iterative optimization shown above, and the cluster centers c_j were update by:

$$p_{ij}=\frac{1}{\sum_{k=1}^{C}\left(\frac{S_{ij}}{S_{ik}}\right)^{\frac{2}{w-1}}} \qquad (6)$$

Every data point essentially belongs to at least one obvious cluster and, therefore, was not eliminated.

$$c_j=\frac{\sum_{i=1}^{N}p_{ij}^{m}(Z_i)}{\sum_{i=1}^{N}p_{ij}^{m}} \qquad (7)$$

The iteration stops when $\max_{ij}\left\{\left|p_{ij}^{k+1}-p_{ij}^{k}\right|\right\}<\eta$, where η is a termination criterion between 0 and 1, and k is the iteration steps. This procedure converges to a local minimum or a saddle point F_{ob}. The FCM process was repeated until it converges. The topic's clusters as of the amassed data were created using this approach. Business or economic effects (work, impact, business, crisis, and pay), anti-discrimination (police, death, protest, die, right, kill, and war), sports (player, football, season, game, league, team, play, and sport), charity (assist, fund, support, relief, million, donate, offer, society, and food), entertainment (show, watch, video, love, and like), politics (government, response, president, state, and minister), case reports, treatment (patient, drug, vaccine, hospital, treatment, hydroxychloroquine, trial, and plasma), online classes (webinar, live, join, talk, impact, discuss, tomorrow, virtual, and host), and additional personal care (wear, mask, people, test, school, social, reopen, and spread) are the 10 disparate clustering topics of the CSPK-FCM's output. The correlated words of the specific topic were represented by the words specified within the bracket of topics.

E. Sentiment Analysis

The SA of clustered data and emoticon features was conducted using the proposed XAI system, like FL-ILSTM after TM. The proposed technique is an integration of 2 algorithms, namely Fuzzy Logic (FL) and improved LSTM (ILSTM). FL is used for precise prediction, and ILSTM is used to explain the prediction. If the data (output) as of the original Decision-Making (DM) model (FL) is Inputted in a New Explainable Model (ILSTM), and the inputs of the DM model (original) are employed as outputs in the explainable model, then the explainable model's outputs shall act as reasoning to an action that occurred throughout the DM prediction process, which is the method's hypothesis. The FL's outputs were utilized as inputs to the ILSTM for proving the hypothesis. After that, the system was trained to get outputs (equivalent inputs from FL) using the ILSTM, which describes why a decision was made in the first place. In other words, the decision has been made due to the ILSTM's outputs. Happy, sad, disgust, anger, surprise, and fear are the 6 basic emotions of the output class in this framework that is detected from the topics. For the SA of COVID-19 data, this sort of FL-centered prediction and the ILSTM-based explanation is called FL with ILSTM (FL-ILSTM). Fig. 2 depicts the framework of FL-ILSTM.

F. Fuzzy Logic

For SA prediction, the clustered data and emoticon features' score value is initially given to FL. DM in humans is imitated by the FL's approach. FL is multi-valued in contrast to two-valued Boolean logic. It comes with degrees of membership along with degrees of truth. A fuzzy rule can be specified as a conditional statement for the sentiment's prediction in topics, and that can well be represented by the subsequent Eqs. (8) and (9).

Rule 1: If q_1 is b_1 and q_2 is d_1, then,

$$R_1 = \omega_1 q_1 + \xi_1 q_2 + \psi_1 \tag{8}$$

Rule 2: If q_1 is b_2 and q_2 is d_2, then,

$$R_2 = \omega_2 q_1 + \xi_2 q_2 + \psi_2 \tag{9}$$

where are the fuzzy sets $q_1 \, and \, q_2$ values are the disparate topic's clustered data attained from CSPK-FCM, and $\omega_1, \xi_1, \psi_1, \omega_2, \xi_2, and\, \psi_2$ values are the parameter set.

1) Improved LSTM

The predicted FL classes are inputted to ILSTM for elucidation after prediction. LSTM is a type of RNN model for sequence prediction. An LSTM unit comprises a cell, an input gate, an Output Gate (OG), and a Forget Gate (FG). Over random time periods, the values are remembered by the cell, and the flow of information into and out of the cell is regulated by the 3 gates. Fig. 2 exhibits the LSTM's structure. An LSTM network maps between an input sequence $k = \{k_1, k_2, \ldots k_T\}$ and an output one $o = \{o_1, o_2, \ldots o_T\}$ by estimating the network unit ctivations using Eqs. (10)–(13). The FG assists in processing the o_{t-1} preceding state output and making decisions by forgetting unwanted information.

Fig. 3 depicts FL-ILSTM for sentiment analysis. The FG with the sigmoid function is defined in Eq. (10). The input gate adds new information with suitable scaling, the values are updated by the sigmoid Activation Function (AF), and the new candidate values are formed by a tanh function (Eqs. (11) and (12)). The updated candidate value (new) with suitable scaling is provided in Eq (13):

$$f_t = S_\sigma\left(w_f . [o_{t-1}, k_t] + b_f\right) \tag{10}$$

$$g_t = S_\sigma\left(w_g . [o_{t-1}, k_t] + b_g\right) \tag{11}$$

$$\tilde{h}_t = \tanh\left(w_h . [o_{t-1}, k_t] + b_h\right) \tag{12}$$

$$h_t = f_t(h_{t-1}) + g_t(\tilde{h}_t) \tag{13}$$

Lastly, the sigmoid function's pertinent output is defined in the subsequent Eqs. (14) and (15).

$$l_t = S_\sigma\left(w_l . [o_{t-1}, k_t] + b_l\right) \tag{14}$$

$$h_t = l_t(\tanh(h_t)) \tag{15}$$

where $g, f, and\, l$ are the input, forget, and OG layers, respectively, h_t is the memory cell state, $\tilde{h}_t$ is the candidate vector, and $w \& b$ in every equation indicates the specific layer's weight and bias values, respectively.

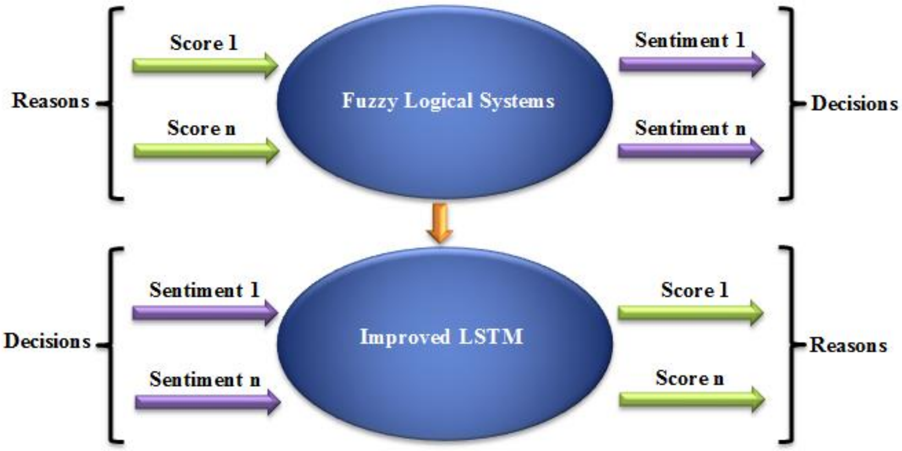


Figure 2. Framework of FL-ILSTM.

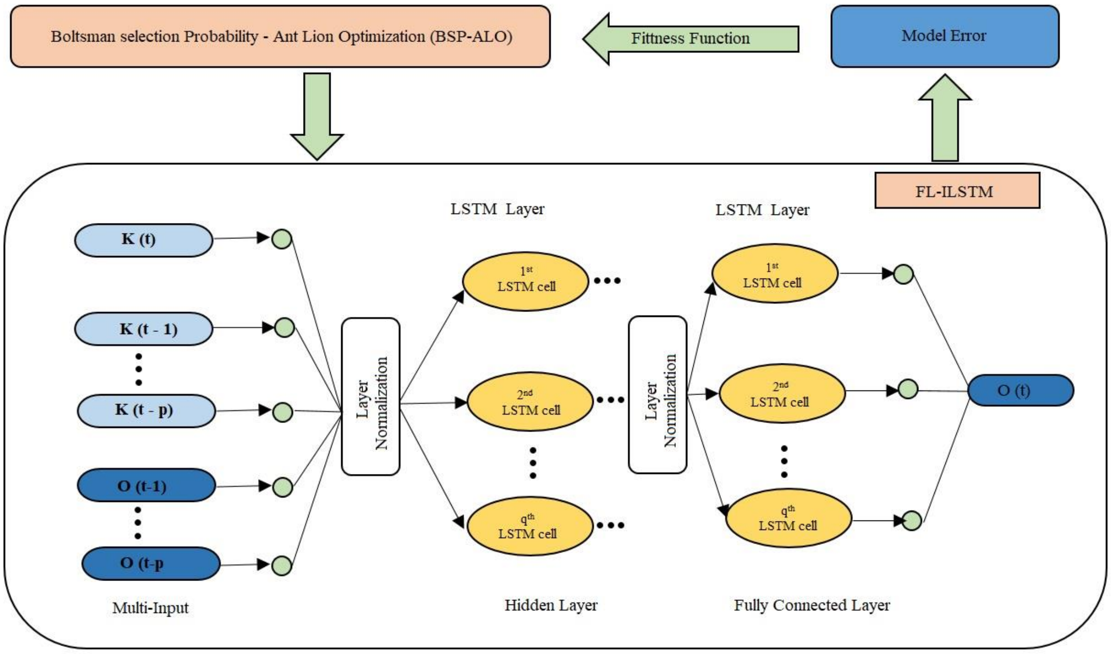


Figure 3. FL-ILSTM for sentiment analysis.

The last layer in the network is the output layer which is utilized for the prediction of sentiment. The later output only comprises the AF aimed at the prediction, and the softmax AF is utilized here. However, a vanishing gradient problem might emerge from the sigmoid activation's usage. Therefore, the vanishing gradient concern was overcome and rectified using a sigmoid function. Besides, ReLU is computationally low-cost analogized to sigmoid since it contains simple mathematical operations. The equation for ReLU is

$$\kappa_R(x) = \max(0, x) \tag{16}$$

Thus, the Eqs. (10)–(14) can be changed by the κ_R activation instead of S_σ. The LSTM network's performance in categorizing disparate problems is promising. Training these networks as other Neural Networks (NN) relies greatly on a variety of hyper-parameters (weights and biases) that decide numerous aspects of algorithm behavior. Therefore, it is essential to optimize LSTM hyperparameters for attaining a successful performance for emotion classification. The Boltzmann Selection Probability based Ant Lion Optimizer (BSP-ALO) for optimizing the LSTM was utilized in this paper. The antlions' hunting behavior was mathematically modeled for optimizing the LSTM network's weights and biases in BSP-ALO. The improved LSTM for the proffered classification system is provided by this performance enhancement in LSTM using ReLU activation and optimization. The BSP-ALO's detailed elucidation is presented below.

2) BSP-basedALO for LSTM optimization

A nature-enthused algorithm that follows the hunting behavior of antlion larvae is called the Ant Lion Optimization (ALO) [32] method. An antlion digs a cone-shaped pit and conceals its larvae below the cone-shaped pit's bottom to trap the ant. The cone's edge is so sharp that the ant could easily fall to the bottom. The antlion throws the sand towards the cone's edge after the prey is trapped in the cone, which makes the prey unable to escape the trap. Then, the antlion consumes the prey and makes a pit for trapping the next prey. The ants seek a Food Source (FS) arbitrarily for foraging at the ALO's initial stage. The algorithm's rapid convergence ability can be strengthened if the arbitrary number selected is larger. Nevertheless, it might delude the algorithm to be trapped in the local optimum. In contrast, a small selection of random values can enhance the algorithm's global-search ability. Nevertheless, the small arbitrary value will cause a slow convergence speed. Boltzmann Selection Probability (BSP) was adapted in the ALO as the base for the ant to choose the FS. Hence, the ALO's selection value can be dynamically modified in the evolution process by utilizing BSP. This BSP Random Walk (RW)-based ALO is called (BSP-ALO). The populace of an arbitrarily generated group of solutions is initialized. For the LSTM network, the probable solutions of BSP-ALO are the weights and biases. The steps in BSP-ALO are defined as follows:

Step 1: The ants, along with the ant lions' local position, are specified with the succeeding matrix Eq. (17).

$$P_{ant} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots a_{1,d} \\ a_{2,1} & a_{2,2} & \cdots a_{1,d} \\ \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots a_{n,d} \end{bmatrix}, P_{al} = \begin{bmatrix} al_{1,1} & al_{1,2} & \cdots al_{1,d} \\ al_{2,1} & al_{2,2} & \cdots al_{1,d} \\ \vdots & \vdots & \vdots \\ al_{n,1} & al_{n,2} & \cdots al_{n,d} \end{bmatrix} \tag{17}$$

where P_{ant} along with P_{al} implies the matrix for storing ants and antlions' position.

Step 2: The fitness of both ants and antlions are signified in matrix form, which is as:

$$F_{ant} = \begin{bmatrix} F(a_{1,1} \;\; a_{1,2} \cdots a_{1,d}) \\ F(a_{2,1} \;\; a_{2,2} \cdots a_{1,d}) \\ \vdots \;\; \vdots \;\; \vdots \\ F(a_{n,1} \;\; a_{n,2} \cdots a_{n,d}) \end{bmatrix}, F_{al} = \begin{bmatrix} F(al_{1,1} \;\; al_{1,2} \cdots al_{1,d}) \\ F(al_{2,1} \;\; al_{2,2} \cdots al_{1,d}) \\ \vdots \;\; \vdots \;\; \vdots \\ F(al_{n,1} \;\; al_{n,2} \cdots al_{n,d}) \end{bmatrix} \quad (18)$$

where F_{ant} and F_{al} are the matrix for storing the Fitness Value (FV) of ants and antlions.

Step 3: The Random Walk (RW) of ants is conducted by utilizing the BSP. The ants are moved in the Search Space (SS) for food and shelter, and the antlions are hunted by utilizing their traps. Herein, an ant relies on a selection probability for selecting one FS. The solution's selection probability s_p is assessed as

$$s_{pi} = \frac{e^{[F_i/t]}}{\sum_{n=1}^{SN/2} e^{[F_n/t]}} \quad (19)$$

$$t = t_I[\Re_m] \quad (20)$$

where F_i implies the x_i solution FV, which is proportional to the nectar quantity of the FS in the position x_i, $SN/2$ is the number of FS which is equal to the total ants, t is the temperature, t_I is the initial temperature which decides the initial selection pressure, $\Re$ is an adjustable parameter which is utilized for controlling the varying speed of selection pressure.

It must be fixed lesser than one along with m implies the iteration number.

Step 4: For keeping the RW inside the SS, the value of Step 3 is normalized using min-max normalization.

$$x_i^m = \left[\frac{\langle x_i^m - r_i \rangle \langle u_i^t - t_i^m \rangle}{\langle s_i^m - r_i \rangle} + t_i \right] \quad (21)$$

where r_i and s_i are the RW's minimum and maximum values of i^{th} variable, respectively, and t_i^m and u_i^t are the minimum and the maximum values of the i^{th} variable at m^{th} iteration, respectively.

Step 5: The roulette wheel approach defines the modeling of the antlion's hunting ability, which looks for the fittest antlion with a larger probability of capturing the prey throughout the optimization method. The RW is affected by the antlion's traps. This assumption can be imposed using Eqs. (22) and (23).

$$t_i^m = al_j^m + t^m \quad (22)$$

$$u_i^m = al_j^m + u^m \quad (23)$$

where t_i^m and u_i^m are the minimum and maximum of every variable at mth iteration, respectively, the t^m and u^m are the minimum and maximum of every variable for the i^{th} ant, respectively, al_j^m is the ant position of chosen ant lion at m^{th} iteration.

Step 6: The antlion throws sand toward the pit's edge by its mouth after recognizing that the ant is inside the trap, which makes the ant trapped within the pit. This behavior of antlions can be characterized in Eq. (24).

$$t^m = \frac{t^m}{10^w \left(\frac{m}{M}\right)}, \; u^m = \frac{u^m}{10^w \left(\frac{m}{M}\right)} \quad (24)$$

where m is the present iteration, M is the total iteration, and w is a constant based on the existing iteration.

Step 7: Antlions capture an ant that attains the pit's bottom after executing Step 6. Then, the equation updates its position to the latest position.

$$al_j^m = a_i^m, \; if \; F(a_i^m) > F(al_j^m) \quad (25)$$

Step 8: Elitism is finally executed to attain the best solution. The best ant lion is attained and stored as elite, which can well be modeled as follows:

$$a_i^m = \frac{r_{al}^m + r_{el}^m}{2} \quad (26)$$

where r_{al}^m is the RW around the chosen antlion, r_{el}^m is the RW near the elite, and a_i^m is i^{th} ant's position.

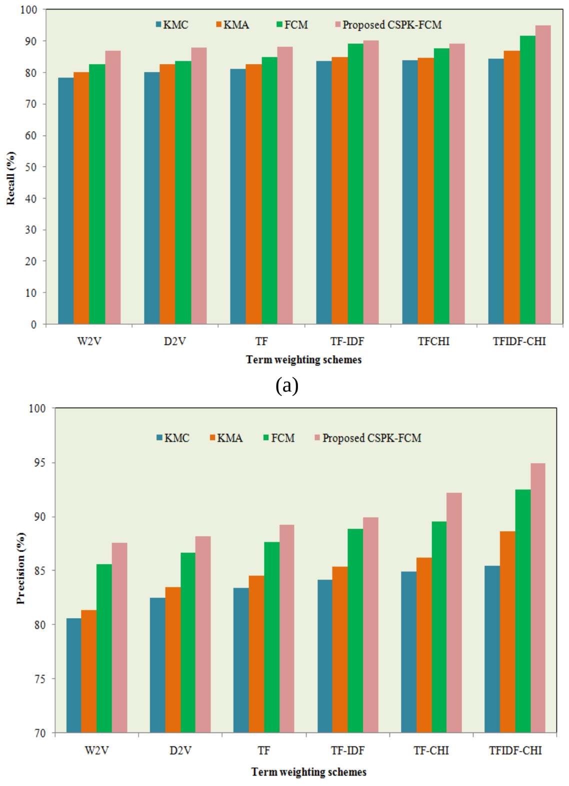
Pseudo code for the BSP-ALO

```
BSP-ALO for LSTM optimization
Input: Weight values of LSTM
Output: Optimized weight values of LSTM
Begin
Initialize the position of $a$ and $al$ in the population
Store the position of $a$ and $al$ in the matrix of $P_{ant}$, and $P_{al}$
Compute the fitness of both $a$ and $al$
Store the fitness values of both $a$ and $al$ into $F_{ant}$ and $F_{al}$
Find the best ant lion and assume it as the $a_i^m$
While the end criterion is not satisfied
        For every ant
                Select a $al$ using Roulette wheel
                Update $t^m$ and $u^m$
                Perform a random walk using BSP, $s_{pi}$
                Normalize $s_{pi}$ using $x_i^m$
                Update the position of ant $al_j^m$, if $F(a_i^m) > F(al_j^m)$
        End for
Compute the fitness of all a
Update $a_i^m$ if an ant lion becomes fitter than the elite
End while
Return elite
```

IV. Result and Discussion

The pseudo-code for the proposed algorithm is given below. Based on sentiments (positive and negative), the tweets are categorized by the classifier and categorized into 8 emotions (fear, joy, anticipation, anger, disgust, sadness, surprise, and trust).

The COVID-19 epidemic that started in 2019 has claimed millions of lives worldwide. The costs associated with treating COVID-19 infection in individuals who have already experienced an acute COVID-19 infection are now readily apparent. The medical research community has extensively used computer analysis of medical images and data during the pandemic. Deep-learning techniques based on Artificial Intelligence (AI) have been applied frequently. The proposed TM and SA model's results are deemed in this section by comparing their efficiency to the existing techniques. The COVID-19 epidemic, initially identified in Wuhan, China, in 2019, has significantly strained the entire world. Schools were closed or transformed to online education to stop the spread of the dangerous virus. Due to the mask requirement, many firms had to close their doors, and other enterprises can follow the release of COVID-19 vaccines. Some people have hesitated to get vaccinated due to rumors, conspiracies, safety concerns, or lockdowns. This study's results showed that positive emotions in various countries are higher when compared to low emotions.

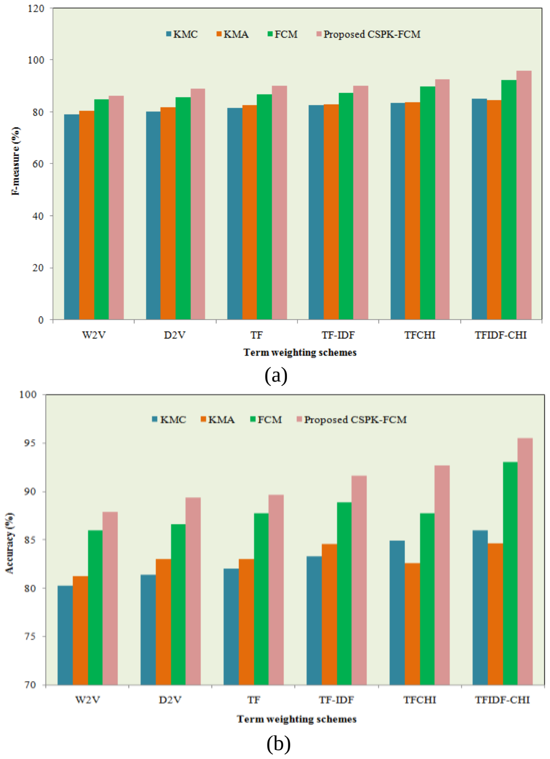


(a)

(b)

Figure 4. Evaluation metrics for clustering models (a) Recall results of clustering models. (b) Precision results of clustering models.

Fig. 4 indicates that the proposed method attains the highest precision and recall for all TW schemes. The proposed CSPK-FCM achieves precisions of 87.62%, 88.23%, 89.28%, 89.99%, 92.23%, and 95.01% for the schemes W2V, D2V, TF, TF-IDF, TF-CHI, and TFIDF-CHI, whereas the existing methods achieve much lower precisions when compared to CSPK-FCM. Comparing the recall attained by the existing classifiers (KMC, KMA, and FCM) for the TFIDF-CHI model is 84.28%, 86.98%, and 91.58%, whereas the recall attained by the CSPK-FCM is 94.87%. It is believed that the proposed clustering model achieves the best precision and recall results compared to other TW schemes. However, combining TFIDF-CHI and CSPK-FCM for topic clustering results in outstanding performance. Fig. 5 depicts the performance comparison with respect to f-measure and accuracy.



(a)

(b)

Figure 5. Evaluation metrics for clustering models (a) F-score results of clustering models. (b) Accuracy result of clustering models.

Utilizing the existing TW schemes, namely W2V, D2V, TF, TFIDF, and TFCHI, the clustering models offer the minimum values of f-score and the highest accuracy, as exhibited by the outcomes. The clustering models achieve the highest level of performance when utilizing the weighted TFIDF-CHI scheme. In contrast, the proposed CSPK-FCM achieves the maximum value of f-score and accuracy for all TW schemes when comparing the results of the existing algorithm with those of the proposed one. Existing models, namely KMC, KMA, and FCM, achieve f-scores of 85.23%, 84.56%, and 92.36% when using the TFIDF-CHI algorithm, whereas the proposed model, CSPK-FCM, achieves f-scores of 95.98% and is similar to others. Similarly, the CSPK-FCM with the TFIDF-CHI TW achieves an accuracy of 95.48%, but the existing KMC, KMA, and FCM's accuracy are 85.96%, 84.58%, and 92.99%, respectively, which are inferior to CSPK-FCM. The highest results of CSPK-FCM with TFIDF-CHI exhibit extraordinary performance in COVID-19 data's

TM. After that, the XAI classification model's results, such as FL-ILSTM, are compared to those of existing classifiers, like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), and LSTM, as listed in Table I. Classifiers are compared in terms of their classification metrics, such as sensitivity, specificity, precision, recall, F-measure, and accuracy, to determine their performance in Section IV.

TABLE I. RESULTS OF CLASSIFIER FOR SENTIMENT ANALYSIS

Performance Metrics (%)	Classifiers					
	SVM	KNN	ANFIS	IKNN	LSTM	Proposed FL-ILSTM
Sensitivity	83.22	85.88	86.56	87.86	90.23	95.89
Specificity	83.88	85.82	86.99	87.96	89.85	95.56
Precision	84.22	85.69	87.89	88.56	90.99	94.32
Recall	85.98	85.75	88.96	88.77	91.26	95.23
F-measure	83.89	83.98	88.65	85.96	90.69	94.58
Accuracy	83.82	84.36	88.26	86.89	91.83	96.88

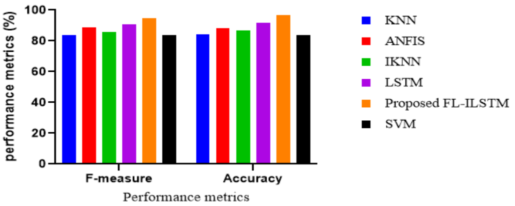


Figure 6. Comparison of classifiers for SA in terms of the f-score and accuracy.

Table I shows the proposed FL-ILSTM's performance efficiency in SC. Compared to the prevalent classifiers, the FL-ILST achieves significantly superior results. The existing work's sensitivity, such as SVM, KNN, ANFIS, IKNN, and LSTM for SC, is 83.22%, 85.88%, 86.56%, and 90.23%, while the FL-ILSTM proposed achieves 95.89%, which is superior to others. Similarly, the proposed FL-ILSTM attains the maximum values for the remaining metrics compared to others. The effective TW and ranking schemes are utilized for scoring the non-emoticon and emoticon features which is the cause for FL-ILSTM-based accurate classification. The similarly discussed topics are clustered using CSPK-FCM to attain the highest accuracy for FL-ILSTM at the topic level. Fig. 6 illustrates the classifier's results regarding f-measure and accuracy.

Table II displays the results of our comparative analysis with other ILSTM models. We found that LE-LSTM and Tree-LSTM models had better accuracy compared to the LSTM model, but their improvements were not significant. It should be noted that Tree-LSTM requires an external tool to parse text and generate a tree structure. Our proposed FL-ILSTM model achieved the best performance on the dataset used in this study. While LE-LSTM had the highest accuracy among all the models, our proposed FL-ILSTM model achieved an accuracy of 96.88%, which is a 3% improvement compared to the LE-LSTM model. Based on these results, we conclude that our proposed methods are effective in improving the accuracy of sentiment analysis. Furthermore, when comparing our proposed methods, we observed that RAE-LSTM and WALE-LSTM did not consistently outperform the FL-ILSTM model. These models only improved the accuracy on datasets with long sequences of text. Fig. 7 illustrates the comparative analysis on other LSTM methodology with proposed FL-ILSTM.

TABLE II. COMPARATIVE RESULTS WITH OTHER ILSTM

Performance Metrics (%)	RAE-LSTM	WALE-LSTM	Tree-LSTM	LE-LSTM	FL-ILSTM
Precision	77.7	89.5	92.3	93.3	94.32
Recall	76.5	80.8	93.3	92.6	95.23
Accuracy	78.9	88.9	92.6	93.6	96.88
F-measure	77.6	88.0	92.4	93.7	94.58

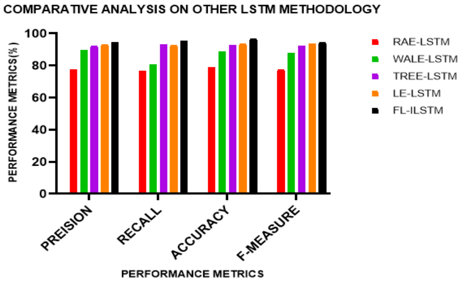


Figure 7. Comparative analysis on other LSTM Methodology.

Table III displays the results of our comparative analysis with other variants of KNN. We found that Fuzzy KNN has better accuracy when compared with other variant models, our proposed model has achieved the best performance on the data set used in this study. Also, our proposed method has proved, the best accuracy when compared with other machine learning algorithms, as well as other variants of ILSTM. Fig. 8 depicts the comparative analysis on other variants of KNN. Our proposed model was also compared with various transformer methods, such as BERT. Fig. 9 showcases their performance in terms of accuracy, precision, and recall. We observed that while the effectiveness of BERT and the proposed FL-ILSTM are relatively similar, our system outperformed other models like CDNN, particularly in terms of accuracy

TABLE III. COMPARATIVE ANALYSIS WITH OTHER VARIANTS OF KNN

Performance Metrics (%)	Hassanat KNN	Mutual KNN	CDNN	ECDNN	Fuzzy KNN	FL-ILSTM
Accuracy	81	77	85	84	83	96.88
Precision	85	89	86	87	89	94.32
Recall	89	83.77	84	87	88	95.23

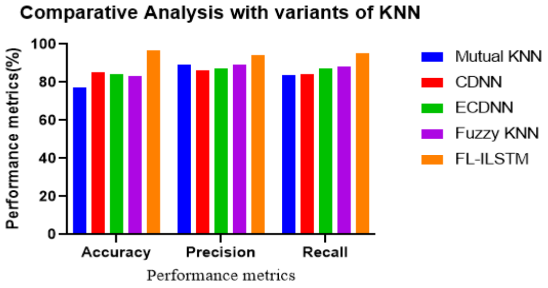


Figure 8. Comparative analysis with other variants of KNN.

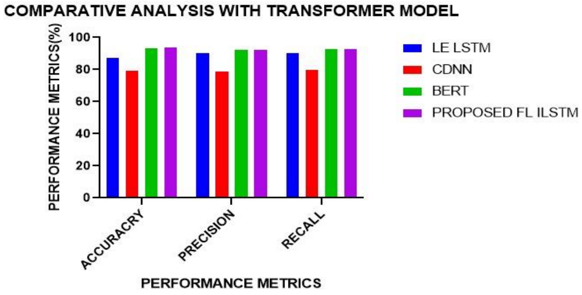


Figure 9. Comparative analysis with transformer model.

The tweet's sentiments and emotions from disparate countries are examined. Table IV displays the positive and negative emotions examined in infected countries. Based on the study's results, most people worldwide have an optimistic outlook. Almost all positive sentiments were represented in the tweets of Australia and Spain, where approximately 62% of the population held positive sentiments, and 42% and 38% had negative sentiments, respectively. 60% of Indian citizens tweeted positively, while 40% negatively perceived it. A balance between positive and negative sentiments is presented in France, Russia, the UK, and South Africa. In contrast, more people in China (54%) tweeted with negative sentiments. This is because China has reportedly reached the curve's peak and is now following a descending trend, whereas other countries are still fighting to prevent the spread and are hopeful.

TABLE IV. SENTIMENT ANALYSIS OF TWEETS ACROSS DIFFERENT COUNTRIES

Infected Countries	Positive (%)	Negative (%)
USA	55	48
Brazil	58	46
South Africa	53	42
Netherland	59	38
China	44	54
India	60	42
UK	56	43
Italy	58	46
Spain	62	44
Russia	50	48
France	52	46
Australia	62	38

Table V displays the examined emotion in countries. It was observed that since the recovery rate is higher in approximately all infected countries, nearly all nations possess the maximum trustworthy tweets. Due to the large number of people infected with COVID-19, most of the initial tweets around the world express fear. There are instances of sadness and disgust around the world. It was believed that the USA, Brazil, France, and China possessed the angriest tweets. The Netherlands and France had the most tweets that expressed sadness. South Africa, the USA, and Brazil have the most tweets expressing fearful emotions. Australia and South Africa had numerous tweets related to Trust and Surprise, while China's tweets revealed the highest level of anticipation. Many tweets exhibited the Joy emotion related to the jokes and memes users shared during the lockdown. UK, Italy, and Brazil had the maximum Joy tweets.

TABLE V. EMOTION ANALYSIS OF COUNTRIES

	Emotions (%)							
Countries	Trust	Sadness	Fear	Anger	Anticipation	Surprise	Disgust	Joy
USA	18.98	12.82	18.23	13.78	9.56	8.97	5.98	11.68
Brazil	16.58	12.39	18.63	12.68	9.69	9.69	7.98	12.36
South Africa	19.52	14.37	20.36	8.95; Y	10.2	9.85	6.52	10.23
Netherland	18.63	18.82	22.58	8.23	12.25	6.32	6.25	6.92
China	19.89	12.66	20.65	12.63	16.25	5.63	5.36	6.93
India	20.33	15.56	19.52	9.21	13.89	6.93	6.31	8.25
UK	20.89	12.42	16.82	7.23	18.25	6.85	4.89	12.65
Italy	21.5	11.82	16.92	7.12	16.22	8.85	5.21	12.36
Spain	20.89	14.45	14.25	8.15	14.23	9.52	6.49	12.02
Russia	19.8	14.92	18.45	10.52	12.5	8.96	6.29	8.56
France	15.1	17.42	21.65	19.23	9.56	6.2	5.82	5.02
Australia	21.56	14.45	20.39	8.25	15.78	6.78	6.47	6.32

A total of 50000 tweets were extracted from the publicly accessible dataset. Kaggle is the publicly utilized dataset for the proposed work, which can be accessed via the following link: https://www.kaggle.com/gpreda/covid19-

tweets. The tweets come under the hashtags namely #covid-19, #COVID19, #CORONAVIRUS, #CORONA, #StayHomeStaySafe, #Stay Home, #Covid_19, #CovidPandemic, #covid19, #Corona Virus, #Lockdown, #Qurantine, #qurantine, #Coronavirus Outbreak, #COVID, and others are compiled for the work. Eighty percent were utilized for training, while the remaining were used for testing. The various machine learning algorithm, variants of LSTM and variants of KNN [33] is being compared with the proposed FL-ILSTM and topic modeling is performed with the base of LDA as in [34] also, utilized Explainable Artificial Intelligence (XAI) [35] for sentiment analysis. In terms of precision, recall, F-measure, and accuracy, the proposed clustering model's results, such as CSPK-FCM, are compared to existing models, including K-Means Clustering (KMC), K-Medoids Algorithm (KMA), and FCM. The model's efficiency is compared to the TW methods, such as Word to Vector (W2V), Document to Vector (D2V), TF, TF-IDF, and TF-CHI. Fig. 4 reveals the clustering model's results in terms of precision and recall.

V. Conclusion

This study executes CSPK-FCM, FL-ILSTM, TM, and SA of COVID-19 Twitter data by proposing these algorithms. Ten disparate topics are clustered for the English tweets, and the SA of these topics is categorized into six fundamental emotions. The technique's performance is evaluated by comparing its results to those of existing ones. As illustrated by the results, the best outcomes for Twitter data are attained by both clustering and classification algorithms. The proposed CSPK-FCM model provides 95.48% accuracy when combined with the TFIDF-CHI scheme. Similarly, the proposed FL-ILSTM provides 96.88% accuracy in SA. Utilizing fuzzy logic systems, FL-ILSTM demonstrates how sentiments (decisions) can be reached. However, the TM and SA implemented in this study are restricted to Twitter users only. Thus, caution is advised before presuming the result's generalizability since everyone in the population does not utilize Twitter. More data must be collected from public tweets to improve sentiment analysis. The daily live streaming of public tweets using related terms will continue. Future classification models will be improved using neural network modeling. The dataset will be evaluated using various ILSTM, and BERT models. The work will be expanded in the coming future by proposing an Explainable Artificial Intelligence (XAI) approach and another deep learning model for executing COVID-19's TM and SA by incorporating the Dravidian language.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, P.C. and D.R.V.P.M.; Methodology, P.C. and D.R.V.P.M.; Software, P.C.; Validation, D.R.V.P.M.; Formal Analysis, P.C. and D.R.V.P.M.; Investigation, P.C. and D.R.V.P.M.; Resources, P.C.; Data Curation, P.C.; Writing-Original Draft Preparation, P.C.; Writing-Review & Editing, P.C. and D.R.V.P.M.; Visualization, P.C.; Supervision, D.R.V.P.M.; all authors had approved the final version.

References

[1] M. Hung *et al*., "Social network analysis of COVID-19 Sentiments: Application of artificial intelligence," *Journal of Medical Internet Research,* vol. 22, no. 8, pp. 1–13, 2020.

[2] A. C Sanders *et al*., "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse," *Medrxiv*, pp. 1–10, 2021.

[3] R. T. Gandhi, J. B. Lynch, and C. D. Rio, "Mild or moderate COVID-19," *New England Journal of Medicine*, vol. 383, no. 18, pp. 1757–1766, 2020.

[4] S. Das and A. Dutta, "Characterizing public emotions and sentiments in COVID-19 environment: A case study of India," *Journal of Human Behavior in the Social Environment*, vol. 31, no. 1–4, pp. 154–167, 2021.

[5] X. Y. Song *et al*., "Classification aware neural topic model for COVID-19 disinformation categorization," *PloS One*, vol. 16, no. 2, pp. 1–22, 2021.

[6] M. Zamani, A. H. Schwartz, J. Eichstaedt, S. C. Guntuku, A. V. Ganesan, S. Clouston, and S. Giorgi, "Understanding weekly covid-19 concerns through dynamic content-specific LDA topic modeling," in *Proc. the Fourth Workshop on Natural Language Processing and Computational Social Science*, 2020, pp. 193–198.

[7] T. D. Melo and C. M. S. Figueiredo, "Comparing news articles and tweets about Covid-19 in Brazil: Sentiment analysis and topic modeling approach," *JMIR Public Health and Surveillance*, vol. 7, no. 2, pp. 1–19, 2021.

[8] K. H Manguri, R. N. Ramadhan, and P. R. M. Amin, "Twitter sentiment analysis on worldwide COVID-19 outbreaks," *Kurdistan Journal of Applied Research*, vol. 5, no. 3, pp. 54–65, 2020.

[9] Z. Shah, D. Surian, A. Dyda, E. Coiera, K. D. Mandl. and A. G. Dunn, "Automatically appraising the credibility of vaccine-related web pages shared on social media: A Twitter surveillance study," *Journal of Medical Internet Research*, vol. 21, no. 11, pp. 1–14, 2019.

[10] Z. Shah and A. G. Dunn, "Event detection on Twitter by mapping unexpected changes in streaming data into a spatiotemporal lattice," *IEEE Transactions on Big Data*, vol. 5, no. 4, pp. 1–16, 2019.

[11] A. Kruspe, M. Häberle, I. Kuhn, and X. X. Zhu, "Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic," arXiv Preprint, arXiv:2008.12172, 2020.

[12] H. J. Jang, E. Rempel, D. Roth, G. Carenini, and N. Z. Janjua, "Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis," *Journal of Medical Internet Research*, vol. 23, no. 2, pp. 1–12, 2021.

[13] P. Resnik, K. E. Goodman, and M. Moran, "Developing a curated topic model for COVID-19 medical research literature," in *Proc. 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020, pp. 1–6.

[14] Q. Liu *et al*., "Health communication through news media during the early stage of the COVID-19 outbreak in China: Digital topic modeling approach," *Journal of Medical Internet Research*, vol. 22, no. 4, pp. 1–12, 2020.

[15] N. A. Deraman *et al*., "A social media mining using topic modeling and sentiment analysis on tourism in Malaysia during COVID-19," in *Proc. IOP Conference Series: Earth and Environmental Science*, 2021, vol. 704, no. 1, pp. 1–10.

[16] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *Journal of Information and Telecommunication*, vol. 5, no. 1, pp. 1–16, 2021.

[17] M. Alam, F. Abid, C. Guangpei, and L. V. Yunrong, "Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications," *Computer Communications*, vol. 154, pp. 1–28, 2020.

[18] Y. K. Gao, Z. D. Xie, and D. M. Li, "Electronic cigarette users perspective on the covid-19 pandemic: Observational study using

twitter data," *JMIR Public Health and Surveillance*, vol. 7, no. 1 pp. 1–7, 2021.

[19] A. A. Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study," *Journal of Medical Internet Research*, vol. 22, no. 4, pp. 1–9, 2020.

[20] T. Mackey *et al*., "Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: Retrospective big data infoveillance study," *JMIR Public Health and Surveillance*, vol. 6, no. 2, pp. 1–9, 2020.

[21] T. Y. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 sensing: Negative sentiment analysis on social media in China via bert model," *IEEE Access*, vol. 8, pp. 38162−138169, 2020.

[22] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets," *IEEE Access*, vol. 8, pp. 181074−181090, 2020.

[23] H. Jelodar, Y. L. Wang, R. Orji, and S. C. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 1–12, 2020.

[24] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of covid-19 tweets using topic modeling, UMAP, and digraphs," arXiv Preprint, arXiv:2005.03082, 2020.

[25] J. Xue *et al*., "Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter," *PloS One*, vol. 15, no. 9, pp. 1–12, 2020.

[26] P. Chakriswaran *et al.*, "Emotion ai-driven sentiment analysis: A survey, future research directions, and open issues," *Applied Sciences*, vol. 9, no. 24, 5462, 2019.

[27] Y. Ren *et al*., "TBSM: A traffic burst-sensitive model for short-term prediction under special events," *Knowledge-Based Systems*, vol. 240, 108120, 2022.

[28] E. K. W. Leow *et al*., "Robo-advisor using genetic algorithm and BERT sentiments from tweets for hybrid portfolio optimization," *Expert Systems with Applications*, vol. 179, 115060, 2021.

[29] C. Huang *et al*., "Sentiment evolution with interaction levels in blended learning environments: Using learning analytics and epistemic network analysis," *Australasian Journal of Educational Technology*, vol. 37, no. 2, pp. 81–95, 2021.

[30] A. X. Wang *et al*., "Implementation and analysis of centroid displacement-Based k-nearest neighbors," in *Proc. the 18th International Conference Advanced Data Mining and Applications*, 2022, pp. 431−443.

[31] A. X. Wang *et al*., "Ensemble k-nearest neighbors based on centroid displacement," *Information Sciences*, vol. 629, pp. 313–323, 2023.

[32] H. Chen *et al*., "A chaotic antlion optimization algorithm for text feature selection," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, p. 41, 2022.

[33] B. P. Nguyen *et al*., "Robust biometric recognition from palm depth images for gloved hands," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 6, pp. 799–804, 2022.

[34] P. Chakriswaran *et al*., "Ensemble of artificial intelligence techniques for bacterial Antimicrobial Resistance (AMR) estimation using topic modeling and similarity measure," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 30, no. 3, pp. 541–565, 2022.

[35] C. Priya and P. D. R. Vincent, "CNS: Hybrid explainable artificial intelligence-based sentiment analysis on COVID-19 lockdown using twitter data," *International Journal of Cooperative Information Systems*, vol. 31, no. 03–04, 2250005, 2023.