Multimedia Tools and Applications (2023) 82:16839–16859
https://doi.org/10.1007/s11042-022-13937-2

Ensemble hybrid model for Hindi COVID-19 text classification with metaheuristic optimization algorithm

Vipin Jain[1] · **Kanchan Lata Kashyap**[1]

Received: 24 March 2022 / Revised: 8 August 2022 / Accepted: 12 September 2022 /
Published online: 24 October 2022

Abstract
A SARS-CoV-2 virus has spread around the globe since March 2020. Millions of people infected worldwide with coronavirus. People from every country expressed their sentiments about coronavirus on social media. The aim of this work is to determine the general public opinion of Indian Twitter users about coronavirus. The Hindi tweets posted about COVID-19 is used as input data for sentiment analysis. The natural language processing is applied on input data for feature extraction. Further, the optimal features are selected from the pre-processed data using the metaheuristic based Grey wolf optimization technique. Finally, a hybrid of convolution neural network(CNN) and a long short-term memory (LSTM) model pair is employed to categorize the sentiments as positive, negative, and neutral. The outcome of the proposed model is compared with other machine learning techniques, namely, Random Forest, Decision Tree, K-Nearest Neighbor, Naive Bayes, Support vector machine (SVM), CNN, LSTM, LSTM–CNN, and CNN–LSTM. The highest accuracy of 87.75%, 88.41%, 87.89%, 85.54%, 89.11%, 91.46%, 88.72%, 91.54%, and 92.34% is obtained by Random Forest, Decision Tree, K-Nearest Neighbor, Naive Bayes, SVM, CNN, LSTM, LSTM–CNN, and CNN–LSTM, respectively. The proposed ensemble hybrid model gives the highest 95.54%, 91.44%, 89.63%, and 90.87% classification accuracy, precision, recall, and F-score, respectively.

Keywords COVID-19 · Sentiment · Grey wolf · Optimization · Deep learning · Ensemble learning

1 Introduction

The coronavirus, also known as COVID-19, has generated a significant public health issue in the past two years. Coronavirus affected the lives of millions of people. Millions of infections and deaths have been reported since March 2020 [27]. The World Health Organization

✉ Vipin Jain
vipin.jain2020@vitbhopal.ac.in

Kanchan Lata Kashyap
kanchan.k@vitbhopal.ac.in

[1] SCSE, VIT University Bhopal, 466114, Madhya Pradesh, India

(WHO) reported 37,109,851 and 1,070,355 confirmed COVID-19 and death cases, respectively [43]. Due to the severity and hazards of COVID-19, the WHO issued emergency guidelines for all medical and public health system on 28 February 2020 [11]. Many people expressed their opinions about COVID-19 on social media such as YouTube, Twitter, and Facebook. The posted messages become the discussion topic among friends and family members with positive or negative response [25]. Social media information affects the life of people positively or negatively [13]. YouTube, Twitter, and Facebook are significant sources of "social data". Sentiment analysis plays a vital role to understand the human emotions by analyzing the human's behaviour [22]. Nowadays, various authors applied the different techniques for the sentiment analysis and classification on social media text. However, a substantial quantity of data about COVID-19 is also available on social media. Historical social data can be utilized by the researchers to give better judgment and conclusion about coronavirus [32]. In this work, Coronavirus-related Hindi tweets posted between 15 March 2020 to 24 May 2021, are collected for sentiment analysis and classification. Many authors have shown the advantages of machine learning techniques for feature extraction and sentiment classification [3, 4, 14, 37]. The combination of lexicon and deep learning techniques can be used to analyze the human sentiments [2].

Motivation The sentiment analysis of Hindi tweets about COVID-19 is done by few authors. In this work, sentiment analysis is performed to understand the Indian sentiments about COVID-19 using Hindi tweets. The various techniques of natural language processing (NLP) with meta-heuristic optimization and a hybrid deep learning model are utilized to analyze the public opinion.

1.1 Contribution of this work

The main contribution of present work is given as:

- A Hindi COVID-19 annotated dataset is developed from Twitter data for sentiment analysis.
- A unique Hindi stop word dictionary is prepared and applied for the experimental analysis.
- The data pre-processing is performed by utilizing various techniques of NLP prior to the construction of labelled feature vectors.
- The Grey wolf optimization algorithm is employed for selection of optimal features.
- Finally, An ensemble deep learning model of convolution neural network (CNN) and long short-term memory (LSTM) is applied for sentiment classification as positive, neutral, or negative.

Organization of the paper The remaining part of the paper is organized as: Section 2 discusses the existing work done by various authors. The proposed methodology is presented in Section 3. The outcome of the proposed model is discussed in Section 4, followed by conclusions in Section 5.

2 Literature review

Many authors investigated the sentiment analysis about COVID-19 tweets recently.

Singh et al. proposed a framework named as Covhindia for sentiment analysis [40]. It detects the emotional polarity of COVID-19 Hindi tweets posted between 23 March and 15 July 2020. The Covhindia framework achieved 88.9% accuracy. Chintalapud et al. investigated 3,090 Indian COVID-19 tweets during the lockdown period [12]. Authors have utilized the Bidirectional Encoder Representations from Transformers (BERT) model for data analysis and achieved 89% accuracy.

The long short-term memory model has been used by Chandra et al. for sentiment analysis [9]. Total 150,000 Indian COVID-19 tweets from March 2019 to September 2020 are examined by the authors. Global Twitter patterns about COVID-19 are examined by Lwin et al. [28]. A lexical approach with the 'Crystal-Feel' algorithm is utilized for text analysis in their work and classified the sentiment into four categories such as joy, sorrow, anger, and fear. It is concluded that the impact of negative feelings dominated the public mental health. Raamkumar et al. analyzed the public attitudes and responses about COVID-19 through social media channels [35]. The public information distribution processes is enhanced by the authors. Zhao et al. analyzed public sentiments by collecting public attention about COVID-19 events at two-month intervals using Sina Microblog of China [44]. Feelings of Indian citizens during the lockdown period is investigated by Barkur et al. [6]. They analyzed both negative and positive sentiments by using 2,400 tweets. Friendly and disputed Twitter keywords of COVID-19 is evaluated by Chen et al. [10]. A dictionary-based language tool is deployed for sentiment analysis in their work. They achieved highest F1-score of 0.9521% by using XLNet on 500,000 sample data. Jelodar et al. presented the LSTM model for sentiment classification of COVID-19 micro-blogs [21]. Kaur et al. investigated public attitudes towards the COVID- 19 by applying NLP and machine learning techniques on 2,058 tweets [24]. They obtained 24.0%, 32.1%, and 43.9% positive, negative, and neutral tweets, respectively. COVID-19 topic modelling and sentiment analysis is done by Prabhakar et al. by utilizing 18,000 tweets with National Research Council (NRC) sentiment lexicon [33]. Nemes et al. utilized a Recurrent Neural Network model to determine the emotional content of tweets as positive or negative [31]. Samuel et al. used Twitter data to discover public opinion on COVID-19 using Naïve Bayes classifier, Logistic regression, and linear regression techniques and obtained highest accuracy of 74% [38].

Gupta et al. evaluated different machine learning techniques, namely, logistic regression, Naive Bayes approach, Support Vector Machine(SVM), and Decision Tree, for sentiment analysis of Hindi tweets [19]. They have used the National Research Council (NRC) Emotion and Hindi Senti-WordNet Lexicon to identify the emotion of phrases. Finally, an integrated convolutional neural network (CNN) is presented for sentiment analysis of 23,767 Hindi tweets as positive, negative, or neutral with 85% accuracy. A subjective lexicon with a graph-based approach has been developed by Arora et al. for Hindi reviews classification and achieved the highest 74% accuracy [5]. Mittal et al. [30] generated an annotated corpus for the Hindi language and achieved 80% classification accuracy by using HindiSentiWordNet (HSWN). Nevertheless, most of the research work has been done by analyzing social media networks. In the recent study, Basile et al. examined the dramatic occurrence effect of social media post [7]. Rigorous research is conducted by the authors on the Reddit social network based wide range of topics and languages.

3 Framework of proposed methodology

The proposed methodology presents sentiment analysis and classification of COVID-19 Hindi tweets composed of three steps: (1) Data collection of Hindi tweets related to COVID-19,

(ii) Data pre-processing and feature extraction using NLP followed by GWO based feature selection, and (iii) Sentiment classification as positive, negative, or neutral. The flow diagram of the proposed framework is presented in Fig. 1. Step wise summary of the proposed work is given in Algorithm 1. A detailed description of each step is discussed in the next subsections.

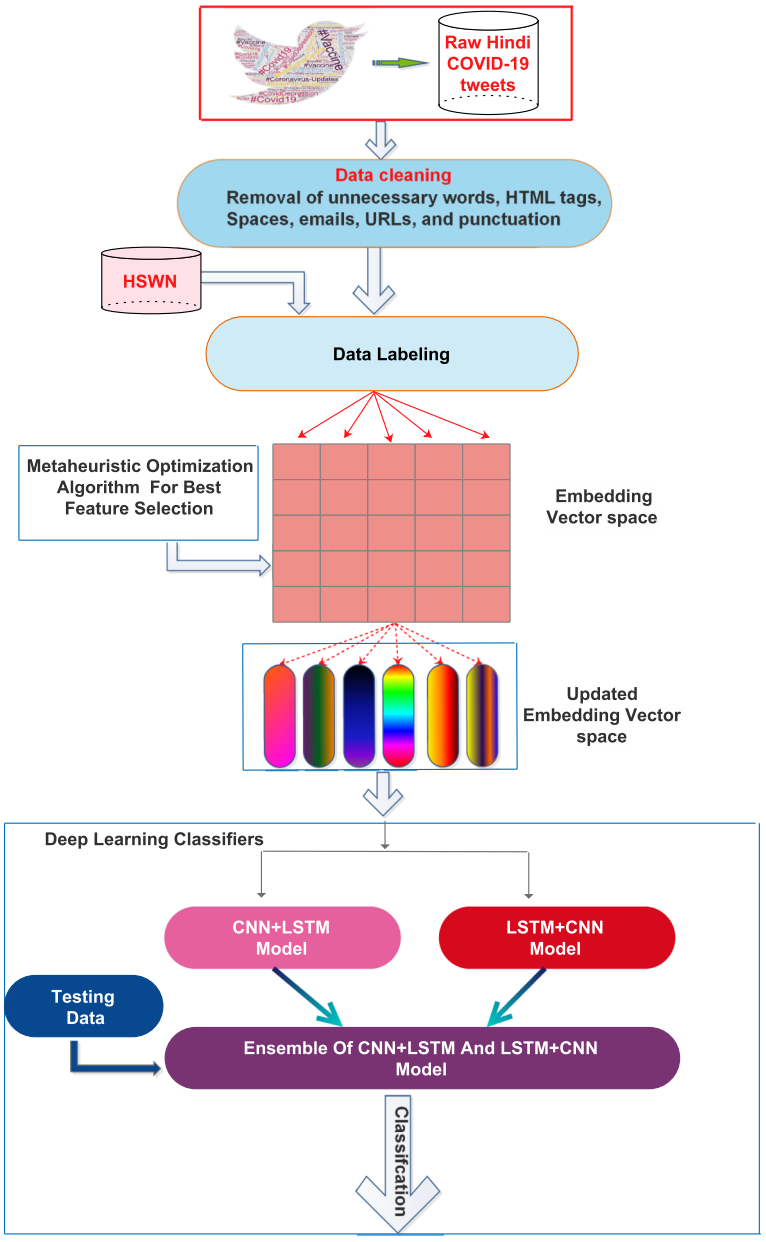

Fig. 1 Overview of the proposed framework

Require: COVID-19 Hindi tweets
Ensure: Sentiment classification as positive, negative, or neutral
1: Scraping of tweets using twint library.
2: Store the scrapped data frame as a dataset.
3: Apply data cleaning and pre-processing on data frame *gh1*
a: *gh1* ⇐Delete the null value.
b: *gh1* ⇐Remove stop words, @, and URL from dataset.
c: *gh1* ⇐Remove emoticons and punctuation from *gh1*.
d: *gh1* ⇐Tokenization.
4: *ZA["qt"]* ⇐Calculate polarity score of tweets in the dataset *gh1*.
5: Assign sentiment to the text based on the polarity score.
6: ***for*** each *ZA[a]*, where a =0,1, 2,3, 4..... n **do**
7: **if** *ZA["qt"]* score > 0 **then**
8: Assign as "Positive"
9: **elseif** *ZA ["qt"]* score < 0 **then**
10: Assign as "Negative"
11: **else**
12: Assign as "Neutral"
13: **end if**
14: ***end for***
15: ***ZA["Sentence"]*** ⇐ Divide all emotions assigned in above steps into three categories such as (i) positive, (ii) negative, and (iii) neutral.
16: ***For*** each ***ZA["Sentences"]*** **do**
17: The index value of each word is calculated.
18: ***end for***
19: Create a ***weight matrix*** by using (2) for each sentence with indexed value of a token.
20: Evaluate the fitness value of each weight of weight matrix by applying fitness function given in (3) and (4).
21: ***Updated_weight_matrix*** ⇐Fitness_function[***weight matrix***]
22: ***Optimized_weight_matrix*** ⇐GWO[***Updated_weight_matrix***]
23: Train the proposed hybrid deep learning with ***Optimized_weight_matrix***

Algorithm 1 Hindi text sentiment analysis of COVID -19.

3.1 Data collection

Hindi tweets about COVID-19 is used in the present work for sentiment analysis. All tweets are collected from Twitter by using Python twint library with the different search keyword namely, #coronavirus, #indiafightscorona, #stayhome, #staysafe, #coronavirusindia, #lockdownindia, #coronaindia, #coronavirusinindia, #COVIDindia, and #COVID-19 as shown in Fig. 2. The Python is used for implementation of scraping script. Total 9,527,810 COVID-19 Hindi tweets are collected from Twitter of the 15 March 2020 to 24 May 2021. The collected tweets contains unstructured, unlabelled, informal, and noisy text. The collected dataset also contains a large number of text features with enormous density. The sample Hindi tweets used for experiment is shown in Fig. 3.

Hashtags	
#lockdownindia	#stayhome
#coronavirusindia	#coronavirusinindia
#covidindia	#staysafe
#covid-19	#indiafightscorona
#coronaindia	#coronavirus

Fig. 2 Keywords applied for tweets crawling

3.2 Data pre-processing

The pre-processing step is used for (i) data cleaning, transformation, and normalization, (ii) feature extraction, and (iii) feature selection. The pre-processing step is summarized diagrammatically in Fig. 4. This step reduces the memory requirements and accelerates the next data processing task. Each pre-processing steps are discussed in the next subsections.

Data cleaning Unwanted data is removed from the collected tweets data in the cleaning process. Regular expression is used in this step which includes (a) Removal of unnecessary words, (b) Elimination of HTML tags, (c) Elimination of emojis and data numbering patterns, (d) Removal of additional characters in sentences, and (e) Elimination of spaces, emails, URLs, and punctuation.

Removal of stop word First, a unique dictionary of 576 Hindi stop words is generated as shown in Fig. 5. All stop words are eliminated from the cleaned data.

Tokenization The tokenization process partitioned the large chunks of text into single piece of word known as token. The indic NLP library is used for the tokenization process [26].

Data labelling Data labelling process is followed by data cleaning for labelling of unlabelled data. The HindiSentiWordNet (HSWN) is used for data labelling and sentiment

	Hindi-Text
0	कोविड-19 की वजह से खेल अथवा मैच तो ख़त्म हो रहे है, विधायकों एवं सांसदों ने जिम्मेदारी ली है , भारतियों के मनोरंजन की Html:col.cvgf-1025//yt
1	कोरोना कोविड पूंजीपतियों एवं उनके दलालों का एक जबरदस्त अन्तर्राष्ट्रीय षड़यंत्र ही तो है | जो भी हॉस्पिटल गया वह पन्नी में बंद हो कर आया , घर वालों को देकने तक नहीं दिया, पता नहीं किडनी निकाली या कुछ ओर, कोई शक@shank |@#$$&&**^^&*|
2	#कोविड -19 गेंदबाजों के लिए बन रहा है कहर @BCCI.###!!!!!!!!!!!!!!!!
3	@कोविड -19 चिताओं को हवा देते हुए ???????@| - टाइम्स ऑफ़ इंडिया !!!!
4	लग्जीरियस "क्रिप्टो " क्रूज शिप सतोशी, बिकेगा अब कौड़ियों के भाव | 777 कमरों का सपना कोविड के कारण बेकार |
5	देश में 24 घंटे में आये 24578 केस $$%%&&***
6	कोरोना के नए स्ट्रेन को लेकर बड़ी राहत, भारत सरकार ने कहा- UK और दक्षिण अफ्रीका में पाए गए वेरियंट्स पर भी वैक्सीन कारगर----------||||||||||||||||||||||||
7	ब्रिटेन से लौटे 38 लोग कोरोना संक्रमित , सात में वायरस स्ट्रेन मिला : #सत्येंद्र जैन

Fig. 3 Contents of raw dataset

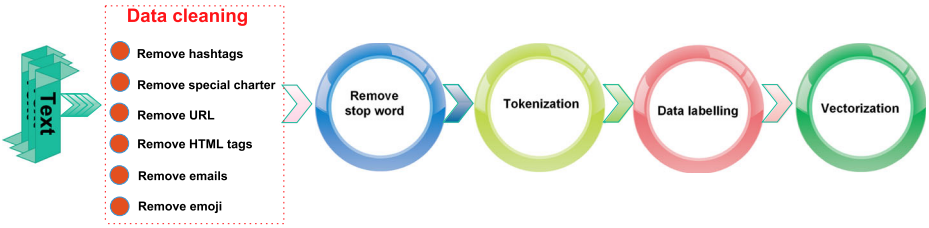

Fig. 4 Pre-processing steps

extraction associated with dataset [23]. This step generates vocabulary which contains Hindi sentimental expressions and their corresponding positive, neutral, and negative polarity.

Vectorization of sentences Vectorization is the final step of pre-processing which maps the pre-processed data into a vector of real numbers. Vectorization is done by indexing technique and mathematically expressed as:

$$S_m = [W_{m,1}, W_{m,2}, \ldots, W_{m,n}] \tag{1}$$

Fig. 5 List of Hindi stop words

Set M as maximum iteration.
Set population B_v (v =1,2,...,q).
Initialize d, E, and G.
Determine the wolves fitness level.
X(α)= Most effective search agent.
X(β)= Second effective search agent.
X(δ)= Third effective search agent.
while Y < M **do**
    for each search agent **do**
        Reposition the active search agent.
    end for
    Update the value of d, E, and G.
    Determine the fitness level of all search agents.
    Update the value of (α), (β), and (δ).
    Y=Y+1
end **while**
return B(α)

Algorithm 2 Algorithm of GWO technique.

here, S_m, m represents the m^{th} sentences and order of the m^{th} sentence, respectively. $W_i,1$, $W_i,2, \ldots, W_i, n$ denotes the weighting vector. The vector length is represented by n. The generated weight matrix is represented mathematically as:

$$weight\_matrix = \begin{bmatrix} W_{(1,1)} & W_{(1,2)} & \cdots & W_{(1,n)} \\ W_{(2,1)} & W_{(2,2)} & \cdots & W_{(2,n)} \\ \vdots & \vdots & \cdots & \vdots \\ W_{(m,1)} & W_{(m,2)} & \cdots & W_{(m,n)} \end{bmatrix} \quad (2)$$

The weight matrix is given as input for the feature selection.

3.3 Grey wolf optimization

Meta-heuristic-based Grey wolf optimization technique (GWO) is used to select the optimal features from the pre-processed data [29]. This technique is used in this work as (i) It is simple and user-friendly and (ii) The convergence rate is faster than other optimization techniques such as Salp swarm algorithm, Firefly, and Harris Hawks. The Updated_weight_matrix is given as initial population to GWO algorithm which generates ***optimized_weight_matrix***. The steps of the GWO algorithm is summarized in Algorithm 2.

Grey wolves represent the Canidae family members and contain strong class structures [15]. Grey wolf always lives in a pack which is a group of five to twelve members [18]. Prey hunting is the most common activity among a pack of wolves. The GWO hierarchy initially covers four ranks of wolves denoted as alpha(α), beta(β), delta(δ), and omega(ω) as shown in Fig. 6.

The α is most powerful wolf among all four groups which takes decisions for various activities such as hunting, discipline, sleep, and get-up time. The second powerful wolf group β supports the most powerful group α for decision-making activities. The delta group

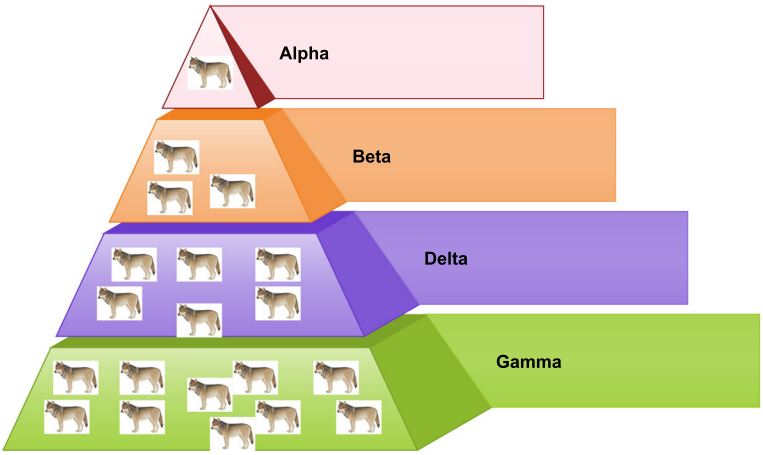

Fig. 6 Social hierarchy of wolves

occupies third place in the wolf social hierarchy. Lastly, ω group ensures security and competency for all wolf packs [1, 39].

3.3.1 Fitness function

The mean absolute difference (MAD) is employed as the fitness function in GWO technique to determine the word relevancy. The fitness function evaluates the word relevancy based on its weight. The ***weight_matrix*** is given as input to the fitness function. It calculates the weight of each word based on its fitness value, and returns the updated weight matrix as output. A relevance score is assigned by the fitness function to each text by comparing their mean value by using a mathematical equation expressed as:

$$MAD(Xh_m) = \frac{1}{Xh_m} \sum_{p=1}^{n} | \, r_m, p - \overline{r_b} \, | \tag{3}$$

$$\overline{r_b} = \frac{1}{Xh_m} \sum_{P=1}^{n} r_{m,P} \tag{4}$$

Here, Xh_m denotes the number of text features taken from the sentences X, r_m represents the mean value of vector m, the weighting value of feature P is denoted by $\overline{r_b}$. n denotes the Total number of text features.

3.4 Mathematical model of Grey wolf algorithm

Gray wolf algorithm incorporates four quantitative methods, namely, (i) Hierarchical structure of Grey wolves, (ii) Prey encircling, (iii) Prey hunting, and (iv) Prey searching and attacking. Each method is briefly discussed in the following subsection.

3.4.1 Hierarchy structure of Grey wolves algorithm

The Grey wolf algorithm focuses on the quantitative hierarchy of the wolf pack leadership. Alpha (α) represents the highest level of intellectual hierarchy, whereas (β) denotes the second and third most essential traits, respectively [20].

3.4.2 Prey encircling

All grey wolves hunt their prey by encircling process which demonstrates the wolves movement in the surroundings. This quantum mechanics can be represented mathematically as [8]:

$$H = \mid G \cdot B_U(y) - B(y) \mid \tag{5}$$

$$B(y+1) = B_U(y) - E \cdot H \tag{6}$$

here, y represents the current iteration, $\mathrm{B}_U(y)$ and $\mathrm{B}(y)$ denotes the position vector of prey and wolf, respectively. H denotes the distance between prey and wolf. E and G represent random vectors which are defined as:

$$E = 2d \cdot u_1 - d \tag{7}$$

$$G = 2 \cdot u_2 \tag{8}$$

here, $u_1$and u_2 denotes random vectors between 0 and 1. These vectors determine the closeness of wolves and prey.

3.4.3 Prey hunting

The whole hunting procedure is lead by Alpha wolf group. The grey wolves follows alpha, beta, and delta groups in search of an ideal hunting spot [17, 41]. The whole hunting search process is mathematically represented as:

$$H_\alpha = \mid G_1 \cdot B_\alpha - B \mid \tag{9}$$

$$H_\beta = \mid G_2 \cdot B_\beta - B \mid \tag{10}$$

$$H_\delta = \mid G_3 \cdot B_\delta - B \mid \tag{11}$$

$$B_1 = B_\alpha - E_1 \cdot H_\alpha \tag{12}$$

$$B_2 = B_\beta - E_2 \cdot H_\beta, \tag{13}$$

$$B_3 = B_\delta - E_3 \cdot H_\delta \tag{14}$$

Updated position of the grey wolf is represented by the following equation:

$$B(y+1) = (B_1 + B_2 + B_3)/3 \tag{15}$$

3.4.4 Prey searching and attacking

In this step, the grey wolves attack the victim and stop their movement. The random value of vector E differentiates the wolf from the prey. The value of vector E is given in the range of $[-d, d]$ and the value of d is computed as:

$$d = 2 - (2 * t/Q_v) \tag{16}$$

As a result, if $E < 1$, the wolf is obligated to exterminate the victim. If $E > 1$, the wolf distinguishes itself from the victim and seeks the fittest prey. The prey searching process is affected by the relative position of all three α, β, and δ wolves. The possible hunting location for a grey wolf is shown in Fig. 7.

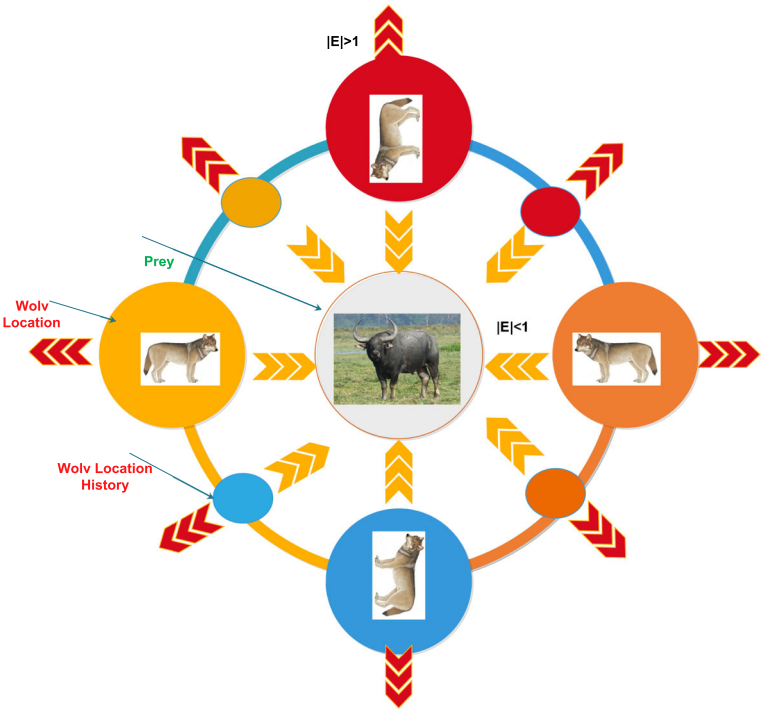

Fig. 7 Wolves hunting movement

The goal of entire process is to determine the calibration of both vectors E and G. The exploitation and exploration are emphasized in almost every dimension. Finally, the GWO algorithm returns an optimal weight matrix denoted as ***optimized_weight_matrix*** and given as input to the classification model.

3.5 Classification model

A hybrid deep learning model of CNN and LSTM is applied for sentiment classification as positive, negative, or neutral. The architecture of the hybrid model is depicted in Fig. 8. The brief description of the CNN and LSTM model is given as:

Convolution neural network model (CNN) CNN model contains convolution layer which minimizes the dimension of input data [16, 34]. Total 50 filters of 3x3 window is used for feature extraction from input data. The extracted features are derived from a set of words by applying the mathematical equation expressed as:

$$M = q(T \cdot h_a + r) \tag{17}$$

Here, M denotes the total number of extracted features, h_a and T denotes the set of words and the filter weights, respectively, r represents a biased factor. The non-linear activation function used by the convolutional layer is given by q.

Long short-term memory model(LSTM) LSTM is a specific recurrent neural network (RNN) model. It uses three gates to maintain and control the long-term dependency along

with the state information of each node. The model can also solve the vanishing gradient problem. The gates and cells of LSTM model is mathematically represented as:

$$\mathrm{P}u_z = \sigma\left(D_{au} \cdot [dp_r - 1], z_r + a_{au}\right) \tag{18}$$

$$\mathrm{E}_r = \tanh\left(D_E\left[dp_r - 1\right], z_r + a_i\right) \tag{19}$$

$$\mathrm{q}_r = \sigma\left(D_q \cdot \left[d_{pq} - 1\right], z_r + a_q\right) \tag{20}$$

$$\mathrm{q}_e = \sigma\left(D_e\left[d_{pr} - 1\right], z_r + a_e\right) \tag{21}$$

here, a denotes the bias vector, D and z_r represents input weight and input vector at time r, respectively. The term P_u, q_e, E_r, and q represents the input gate, output gate, cell memory, and forget, respectively.

Hybrid model Hybrid model is constructed by the hybridization of both CNN and LSTM models. It combines the functionality of convolutional network with the LSTM. The architecture of the hybrid model is shown in Fig. 8. First, embedded words are fed into the CNN-LSTM model through a convolution layer. Output of CNN model is given as input to the LSTM layer for further processing.

Dropout Dropout is an essential trick of deep learning model to prevent over-fitting [36]. It skips the non-participating neurons of the back-propagation process. Dropout strategy removes the neurons to avoid co-adaptation during the training of the model.

Embedding layer This layer embeds the pre-processed dataset with a unique ID and provides a meaningful sequence of words. Index-based weight matrix is employed for word embedding. This layer also assigned a random weight to each word of the training dataset.

Convolution layer The embedding layer transmits sentences to the convolutional layer which uses a pooling layer for convolution with the given input. The pooling layer controls network over-fitting by reducing the input phrases representation.

Global max-pooling Finally, global maximum pooling (max-pooling) is implemented to obtain the best global results.

Activation function The rectified linear unit (RELU) is utilized as an activation function. It returns the output based on the value of input neurons. The mathematical expression of RELU function is given as:

$$R(W) = \left\{ \begin{array}{ll} W & W > 0 \\ 0 & W <= 0 \end{array} \right\} \tag{22}$$

here, W denotes the input neuron.

Dense layer This layer is also referred to as a linked layer in which every neuron is connected with every subsequent neuron. This layer is used to classify input features obtained from the convolution layers.

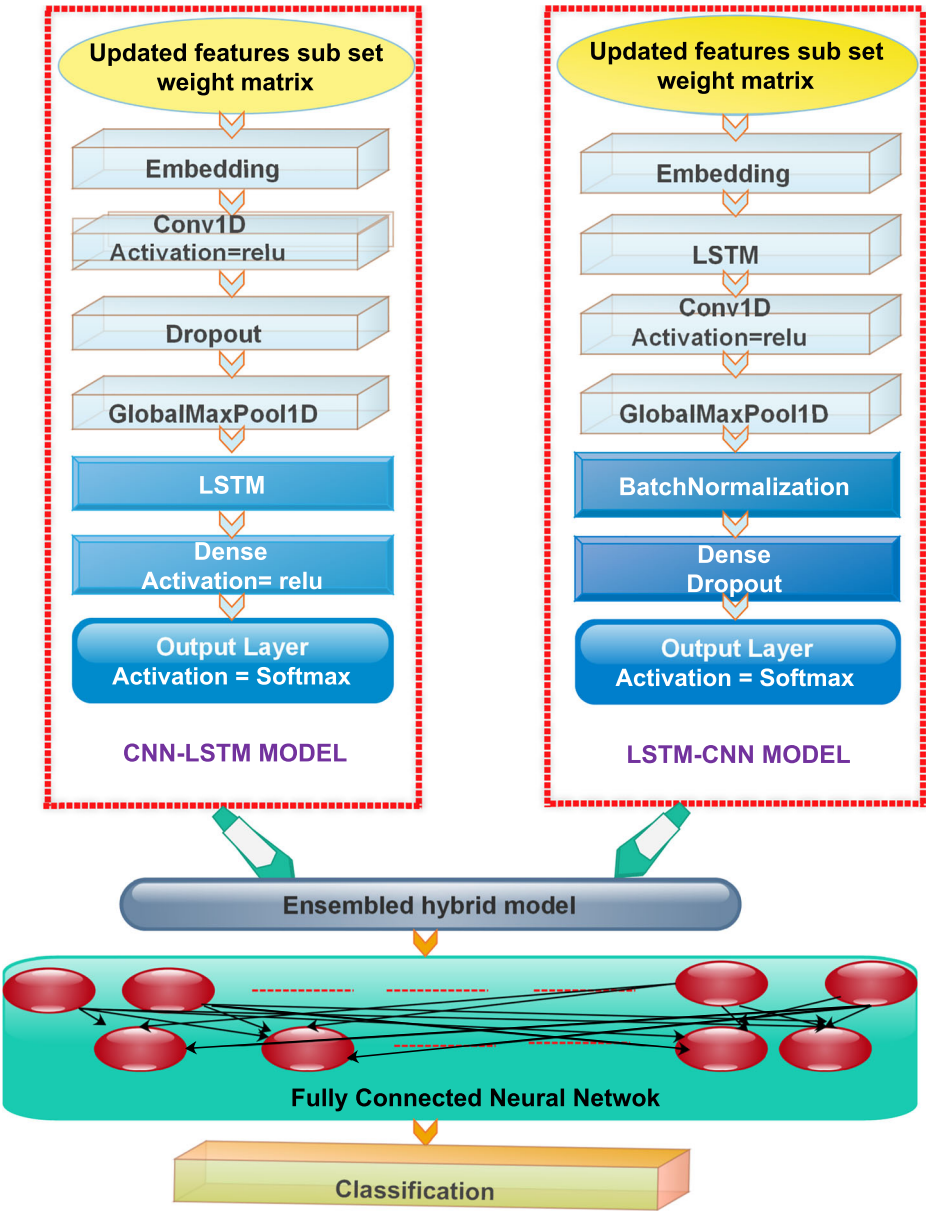

Fig. 8 Architecture of hybrid proposed model

Soft-Max The Soft-Max function is utilized as the last layer of the hybrid neural network. It is mathematically expressed as:

$$\text{soft-max}\left(S_p\right) = \frac{\exp\left(S_p\right)}{\sum_q \exp\left(S_q\right)} \tag{23}$$

here, S represents the output produced by the last layer. The exponential function serves as the non-linear function. The output is normalized by the sum of exponential values and converted into probabilities.

4 Experimental results and discussions

The proposed algorithm is implemented by using Python 3.7 with an i8 E-2236 processor, 32 GB RAM, and NVIDIA P2200 display card.

4.1 Preprocessing results

Total 9,527,810 COVID-19 Hindi tweets from 15 March 2020 to 24 May 2021 are collected. Total 21,914 sarcastic, non-Hindi, and non-subjective tweets are removed from the original dataset. The remaining 9,505,896 tweets is given as input to the proposed framework for further processing. Sample of original text data and extracted tokens are given as:

Sample of original text data

Hindi-Text	English- Translation
कोविड-19 की वजह से खेल अथवा मैच तो ख़त्म हो रहे है, विधायकों एवं सांसदों ने जिम्मेदारी ली है, भारतियों के मनोरंजन की.	Due to Covid-19, sports or matches are ending, MLAs and MPs have taken responsibility, for the entertainment of Indians.

Sample example of extracted tokens

Hindi-Text	English- Translation
कोविड-19, की, वजह, से, खेल, अथवा, मैच, तो, ख़त्म, हो, रहे, है, विधायकों, एवं, सांसदों, ने, जिम्मेदारी, ली, है, भारतियों, के, मनोरंजन की.	Due, to, Covid-19, sports, or, matches, are, ending, MLAs, and, MPs, have, taken, responsibility, for, the, entertainment, of, Indians.

	Hindi-Text	English- Translation	labeled
0	कोविड-19 की वजह से खेल अथवा मैच तो ख़त्म हो रहे है, विधायकों एवं सांसदों ने जिम्मेदारी ली है , भारतियों के मनोरंजन की.	Due to Covid-19, sports or matches are ending, MLAs and MPs have taken responsibility, for the entertainment of Indians.	Positive
1	कोरोना कोविड पूंजीपतियों एवं उनके दलालों का एक जबरदस्त अन्तर्राष्ट्रीय षड़यंत्र ही तो है | जो भी हॉस्पिटल गया वह पन्नी में बंद हो कर आया , घर वालों को देकने तक नहीं दिया, पता नहीं किडनी निकाली या कुछ ओर , कोई शक |	Corona covid is only a tremendous international conspiracy of capitalists and their brokers. Whoever went to the hospital came closed in foil, did not even give it to the family members, I do not know whether the kidney was removed or something else, no doubt.	Negative
2	कोविड -19 गेंदबाजों के लिए बन रहा है कहर .	covid-19 is creating havoc for bowlers	Negative
3	कोविड -19 चिताओं को हवा देते हुए |- टाइम्स ऑफ़ इंडिया	Fueling Covid-19 concerns - Times of India	Neutral
4	लग्जीरियस "क्रिप्टो " क्रूज शिप सतोशी, बिकेगा अब कौड़ियों के भाव | 777 कमरों का सपना कोविड के कारण बेकार |	Luxury "Crypto" cruise ship Satoshi, will now be sold for a penny price. The dream of 777 rooms wasted due to covid.	Negative
5	देश में 24 घंटे में आये 24578 केस	24578 cases came in the country in 24 hours	Neutral
6	कोरोना के नए स्ट्रेन को लेकर बड़ी राहत, भारत सरकार ने कहा- UK और दक्षिण अफ्रीका में पाए गए वेरियंट्स पर भी वैक्सीन कारगर	Happy new year, in the beginning of 2020, covid hurts. In the end, the Finance Minister gave good wishes for 2021 that it will get rid of it soon.	positive
7	ब्रिटेन से लौटे 38 लोग कोरोना संक्रमित , सात में वायरस स्ट्रेन मिला : सत्येंद्र जैन	38 people returned from Britain corona infected, virus strain found in seven: Satyendra Jain.	Negative

Fig. 9 Labeled data

Table 1 Result of sentiment analysis

Total number of tweets	Negative	Positive	Neutral
9,505,896	4,763,905	3,811,124	930,867

4.2 Results of data labelling and sentiment analysis

A sample of labelled text is shown in Fig. 9. The distribution of COVID-19 Hindi tweets based on outcome of sentiment analysis is shown in Table 1. Total 4,763,905, 3,811,124, and 930,867 tweets are analyzed as negative, positive, and neutral, respectively.

4.3 Classification results and discussion

A deep learning-based hybrid model is used for sentiment classification as positive, negative, and neutral. Three different hybrid models namely, (a) CNN-LSTM, (b) LSTM–CNN, and (c) CNN-LSTM + LSTM–CNN are applied for the classification. Performance of the proposed hybrid model is evaluated in terms of accuracy, precision, recall, and F-score. Definition of evaluation parameters are given as:

$$Accuracy = \frac{\sum(G\_D, G\_Z)}{\sum(G\_D, P\_D, G\_Z, P\_Z)} \tag{24}$$

here, G_D, G_Z, P_D, P_Z denote the properly identified, wrongly identified, rejected in the proper manner, and wrongly rejected, respectively. Precision defines the number of data tagged as positive by machine learning models.

$$Precision = \frac{(G\_D)}{\sum(G\_D, P\_D)} \tag{25}$$

Recall is also referred to as sensitivity or positive predictive value. Mathematically it is defined as:

$$Recall = \frac{(G\_D)}{\sum(G\_Z, P\_Z)} \tag{26}$$

Table 2 Parameter values set as input for all three hybrid models

Parameter name	CNN-LSTM Value	LSTM-CNN Value	CNN-LSTM+LSTM-CNN Value
Epoch	175	143	143
Batch size	264	232	264
Max-pooling layer size	2	2	2
Activation function	Relu	–	Relu
Pooling layer padding	Same	–	Same
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001
Filters	–	64	64
Kernel size	–	5	5

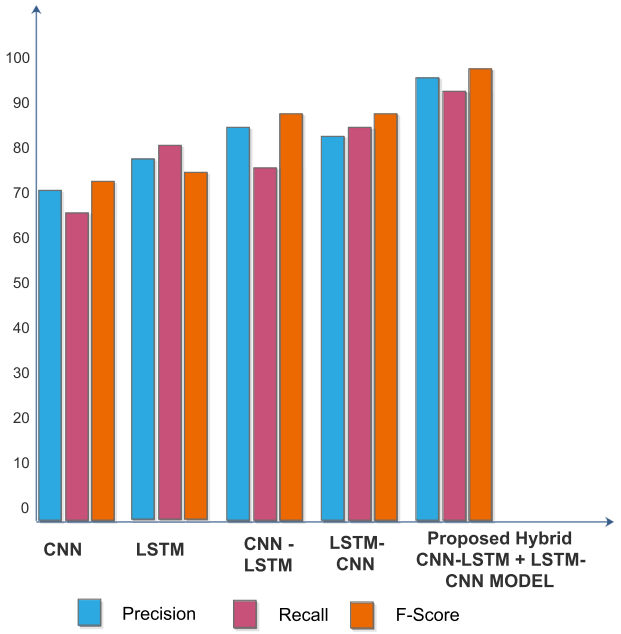

Fig. 10 Comparison of the proposed ensemble hybrid model with CNN, LSTM, CNN-LSTM, and LSTM-CNN

F-Scores is defined the harmonic mean of precision and recall and denoted as [42]:

$$F - score = \frac{2 * (Precision * Recall)}{\sum(Precision, Recall)} \tag{27}$$

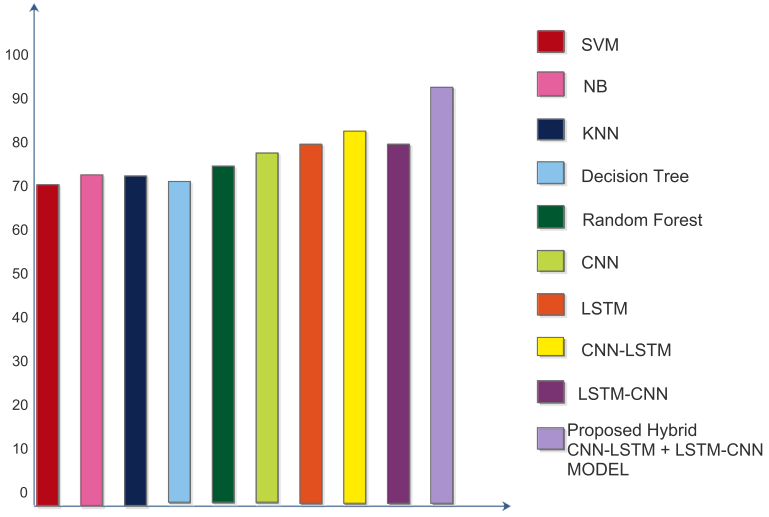

Fig. 11 Comparison of the proposed hybrid deep learning model with standard machine and deep learning model in terms of accuracy

Table 3 Comparison of the proposed model with traditional machine and deep learning models

Model	Precision (%)	F-score (%)	Recall (%)	Accuracy (%)
SVM	88.14	86.32	85.63	89.11
NB	89.85	86.89	83.65	85.54
KNN	85.32	78.57	89.36	87.89
DT	86.85	84.20	88.74	88.41
RF	90.12	89.07	91.47	87.75
CNN	87.45	90.25	91.63	91.46
LSTM	88.65	89.785	90.47	88.72
CNN-LSTM	90.23	89.14	90.32	92.34
LSTM-CNN	89.62	87.51	85.32	91.54
Proposed model (CNN-LSTM+LSTM-CNN)	91.44	90.87	89.63	95.54

The various parameter values set as input for all three hybrid models are presented in Table 2.

All classification models are fine-tuned to obtain the best value of each parameter. Vector dimension of the embedding layer is set as 300. Total 64 neurons are used for the hidden layer of the LSTM-CNN+CNN-LSTM model. The learning rate is fixed as 0.001 with the Adam optimizer. A dropout layer is added with a value of 0.5 between the hidden and output layer of both LSTM and CNN models. A fully linked layer of 32 neurons is used to avoid over-fitting. The outcome of CNN and LSTM models are aggregated with distinct hyper-parameters values (Figs. 10 and 11).

The performance of all three hybrid models are compared with traditional machine learning and deep learning models namely, Random Forest(RF), Decision Tree, K nearest neighbor (KNN), Naive Bayes (NB), SVM, LSTM, and CNN.

4.4 Comparison of proposed work with existing work

Performance of the proposed model is compared with the existing machine and deep learning models as shown in Tables 3 and 4. The highest accuracy of 89.11%, 85.54%, 87.89%, 88.41%, 87.75%, 91.46%, 88.72%, 82.34%, 91.54%, and 95.54% is achieved with SVM, NB, KNN, DT, RF, CNN, LSTM, CNN–LSTM, LSTM–CNN, and ensemble hybrid model (LSTM-CNN+CNN-LSTM), respectively. It is observed from the result that the performance of the ensemble hybrid model (LSTM-CNN+CNN-LSTM) is better than traditional machine learning models.

Convergence plot and boxplot of GWO are compared with three other optimization techniques namely, Salp swarm algorithm, Firefly, and Harris Hawks, as shown in Figs. 12 and 13, respectively. The X and Y axis of the convergence plot represent the number of

Table 4 Comparison of proposed model with existing work

Model	Accuracy
Purva et al. [40].	88.9%
Chintalapud et al. [12].	89%
Proposed model (CNN-LSTM+LSTM-CNN)	95.54

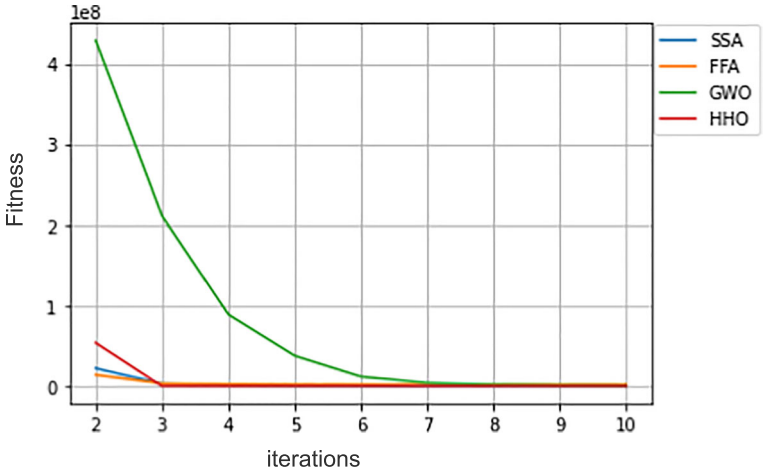

Fig. 12 Comparison of the convergence rates of the GWO and Salp swarm algorithm, Firefly, and Harris Hawks optimization algorithm

iterations and fitness values, respectively. It is analyzed from the convergence plot that the suggested technique converges faster than the Salp swarm algorithm, Firefly, and Harris Hawks optimization technique.

The X-axis and Y-axis of the boxplot represent the applied techniques and fitness value, respectively. The boxplot shows that the proposed methodology gives a better result for all parameter values.

4.5 Analysis of computational complexity

The computational complexity of the entire prediction process is analyzed as:

Data-cleaning process requires a total of O(Twitter posts $\times$ total word count). Feature extraction step requires $T(f) = O(f^2)$ + parsing time. Here, f is the total number of tweets in the sample. Computational complexities of GWO algorithm for feature selection is summarized as:

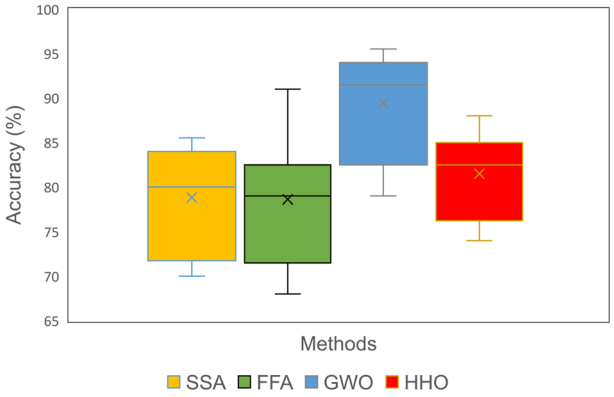

Fig. 13 Comparison of GWO with Salp swarm algorithm, Firefly, and Harris Hawks optimization algorithm

Table 5 Computational complexity of proposed model

Steps	Time-complexity
Data-cleaning process	O(Twitter posts × total word count)
Feature Extraction	$T(f) = O(f^2)$ + parsing time
Features selection with GWO	$O(W \times q \times \text{MI})$
Evolution of sentiments	$O(b \times \text{s}(ac + cx + xy))$
Forecasting	$O(1)$

(i) GWO initialization requires $O(W \times q)$ time, here W and q denote population density and problem dimension, respectively. (ii) Calculation of GWO process parameters require $O(W \times q)$. (iii) $O(W \times q)$ time is required to update the wolf position. (iv) Evaluation of fitness value requires $O(W \times q)$ time. The entire time complexity of the GWO is denoted as $O(W \times q \times MI)$ and here, MI denotes the optimum iterations. The computational complexity of supervised learning algorithms depends on the number of iterations or the number of classification classes. The time complexity of CNN-LSTM model is given as $O(b \times s(ac + cx + xy))$, here, a, c, x, and s denote the number of input layer nodes, second layer node, third layer node, and training examples, respectively. by and y denotes the total number of epochs and output layer nodes, respectively. The prediction step requires a temporal complexity of $O(1)$. The overall time complexity is $O(f^2)$. Time complexity of each stage is shown in Table 5.

5 Conclusions

The meta-heuristic based Grey wolf optimization technique and hybrid deep learning model is presented in this work. Data preprocessing and labelling using natural language processing plays a important step in the proposed model to construct a feature vector. The optimal features are selected by applying the Grey wolf optimization technique. Finally, the sentiment classification has been done by applying a hybrid model of CNN-LSTM and LSTM-CNN. In addition, results of the proposed model are compared with traditional machine learning models also. Proposed model gives highest **95.54%**, 91.44%, 89.63%, and 90.87% of classification accuracy, precision, recall, and F-score, respectively. The experimental results show that the proposed model is more effective than other machine learning models for the sentiment classification of Hindi tweets. This study shows that people are optimistic about COVID-19. Positive tweets suggest that individuals are enthusiastic about COVID-19, while negative tweets indicate that users are scared by its effects. The neutral tweets extracted from the dataset demonstrated that many peoples are confused about the impact of COVID-19. The findings of this study will be beneficial for the government, policymakers, and healthcare management to understand the effects of COVID-19 on society. The findings of this work can be utilized to build more effective strategies for boosting public attitude during the various phases of pandemics. COVID-19 vaccine sentiment analysis will be performed on the Hindi dataset in future work. This will assist the policymakers to address the people concerns prior to mass vaccination. Multi-model framework for sentiment analysis will also be developed by integrating text messages and voice tones about COVID-19 issues.

Funding There is no funding for this research work.

Data Availability All tweets analyzed in this study were collected from Twitter using Python twint library.

Declarations

Ethical approval This article does not contain any studies with human participants performed by any authors.

Competing interests The authors declare no conflict of interest.

References

1. Abualigah LM, Khader AT (2017) Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. J Supercomput 73(11):4773–4795
2. Ahmad M, Aftab S, Ali I (2017) Sentiment analysis of tweets using svm. Int J Comput Appl 177(5):25–29
3. Ambati LS, El-Gayar O (2021) Human activity recognition: a comparison of machine learning approaches. Journal of the Midwest Association for Information Systems (JMWAIS) 2021(1):49
4. Ambati LS, El-Gayar O, El O, Nawar N (2021) Design principles for multiple sclerosis mobile self-management applications: a patient-centric perspective
5. Arora P, Bakliwal A, Varma V (2012) Hindi subjective lexicon generation using wordnet graph traversal. Int J Comput Ling Applic 3(1):25–39
6. Barkur G, Vibha GBK (2020) Sentiment analysis of nationwide lockdown due to covid 19 outbreak: evidence from India. Asian J Psych 51:102089
7. Basile V, Cauteruccio F, Terracina G (2021) How dramatic events can affect emotionality in social posting: the impact of covid-19 on reddit. Future Internet 13:2. https://doi.org/10.3390/fi13020029
8. Bohat VK, Arya KV, Rajput SS (2018) Prey phase based grey wolf optimizer. In: 2018 Conference on Information and Communication Technology (CICT). IEEE, pp 1–5
9. Chandra R, Krishna A (2021) Covid-19 sentiment analysis via deep learning during the rise of novel cases. Plos one 16(8):e0255615
10. Chen L, Lyu H, Yang T, Wang Y, Luo J (2020) In the eyes of the beholder: analyzing social media use of neutral and controversial terms for covid-19. arXiv:2004.10225
11. Cheng JL, Huang C, Zhang GJ, Liu DW, Li P, Lu CY, Li J (2020) Epidemiological characteristics of novel coronavirus pneumonia in henan. Zhonghua jie he he hu xi za zhi= Zhonghua jiehe he huxi zazhi= Chinese Journal of Tuberculosis and Respiratory Diseases 43:E027–E027
12. Chintalapudi N, Battineni G, Amenta F (2021) Sentimental analysis of covid-19 tweets using deep learning models. Infectious Disease Reports 13(2):329–339
13. Crawford K (2009) Following you: disciplines of listening in social media. Continuum 23(4):525–535
14. El-Gayar OF, Ambati LS, Nawar N (2020) Wearables, artificial intelligence, and the future of healthcare. In: AI and Big Data's Potential for Disruptive Innovation. IGI Global, pp 104–129
15. Faris H, Aljarah I, Al-Betar MA, Mirjalili S (2018) Grey wolf optimizer: a review of recent variants and applications. Neur Comput Applic 30(2):413–435
16. Géron A (2019) Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.
17. Gupta E, Saxena A (2016) Grey wolf optimizer based regulator design for automatic generation control of interconnected power system. Cogent Eng 3(1):1151612
18. Gupta S, Deep K (2018) Cauchy grey wolf optimiser for continuous optimisation problems. J Exper Theor Artif Intell 30(6):1051–1075
19. Gupta V, Jain N, Shubham S, Madan A, Chaudhary A, Xin Q (2021) Toward integrated cnn-based sentiment analysis of tweets for scarce-resource language-hindi. Transactions on Asian and Low-Resource Language Information Processing 20(5):1–23
20. Howe Jr WT, Hinderaker A (2018) "the rule was the rule": new member socialization in rigidly structured totalistic organizations. Atlantic J Commun 26(3):180–195
21. Jelodar H, Wang Y, Orji R, Huang S (2020) Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. IEEE J Biomed Health Inform 24(10):2733–2742

22. Ji X, Chun SA, Geller J (2016) Knowledge-based tweet classification for disease sentiment monitoring. In: Sentiment analysis and ontology engineering. Springer, pp 425–454
23. Joshi A, Balamurali AR, Bhattacharyya P et al (2010) A fall-back strategy for sentiment analysis in hindi: a case study. Proceedings of the 8th ICON
24. Kaur C, Sharma A (2020) Twitter sentiment analysis on coronavirus using textblob. EasyChair
25. Kim K-S, Sin S-CJ, Yoo-Lee EY (2014) Undergraduates' use of social media as information sources. College & Research Libraries 75(4):442–457
26. Kunchukuttan A (2020) The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf. Accessed December 2021
27. Liu R, Han H, Liu F, Lv Z, Wu K, Liu Y, Feng Y, Zhu C (2020) Positive rate of rt-pcr detection of sars-cov-2 infection in 4880 cases from one hospital in wuhan, china, from jan to feb 2020. Clin Chim Acta 505:172–175
28. Lwin MO, Lu J, Sheldenkar A, Schulz PJ, Shin W, Gupta R, Yang Y (2020) Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. JMIR Public Health and Surveillance 6(2):e19447
29. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61
30. Mittal N, Agarwal B, Chouhan G, Bania N, Pareek P (2013) Sentiment analysis of hindi reviews based on negation and discourse relation. In: Proceedings of the 11th workshop on Asian language resources, pp 45–50
31. Nemes L, Kiss A (2021) Social media sentiment analysis based on covid-19. J Inform Telecommun 5(1):1–15
32. Pan X, Ojcius DM, Gao T, Li Z, Pan C, Pan C (2020) Lessons learned from the 2019-ncov epidemic on prevention of future infectious diseases. Microbes and Infection 22(2):86–91
33. Prabhakar Kaila D, Prasad DrAV et al (2020) Informational flow on twitter–corona virus outbreak–topic modelling approach. International Journal of Advanced Research in Engineering and Technology (IJARET) 11:3
34. Qing L, Linhong W, Xuehai D (2019) A novel neural network-based method for medical text classification. Fut Int 11(12):255
35. Raamkumar AS, Tan SG, Wee HL et al (2020) Measuring the outreach efforts of public health authorities and the public response on facebook during the covid-19 pandemic in early 2020: cross-country comparison. J Med Int Res 22(5):e19334
36. Ruangkanokmas P, Achalakul T, Akkarajitsakul K (2016) Deep belief networks with feature selection for sentiment classification. In: 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS). IEEE, pp 9–14
37. Sai Ambati L, El-Gayar OF, Nawar N (2020) Influence of the digital divide and socio-economic factors on prevalence of diabetes
38. Samuel J, Ali GG, Rahman M, Esawi E, Samuel Y et al (2020) Covid-19 public sentiment insights and machine learning for tweets classification. Information 11(6):314
39. Sangaiah AK, Fakhry AE, Abdel-Basset M, El-henawy I (2019) Arabic text clustering using improved clustering algorithms with dimensionality reduction. Clust Comput 22(2):4535–4549
40. Singh P (2020) Covhindia: deep learning framework for sentiment polarity detection of covid-19 tweets in hindi. International Journal on Natural Language Computing (IJNLC), 9
41. Tan Y, Takagi H, Shi Y (2017) Advances in swarm intelligence: 8th international conference, icsi 2017, Fukuoka, Japan, july 27–august 1, 2017, proceedings, part i, vol 10385. Springer
42. Tatbul N, Lee TJ, Zdonik S, Alam M, Gottschlich J (2018) Precision and recall for time series. Advances in neural information processing systems, 31
43. WHO (2020) COVID-19 Situation Report, https://www.who.int/publications/m/item/weekly-epidemiological-update—12-october-2020. Accessed Nov-Dec 2020
44. Zhao Y, Xu H (2020) Chinese public attention to covid-19 epidemic: based on social media. medrxiv. Preprint posted online March 20