New Generation Computing (2022) 40:1165–1202
https://doi.org/10.1007/s00354-022-00182-2

KEAHT: A Knowledge-Enriched Attention-Based Hybrid Transformer Model for Social Sentiment Analysis

Dimple Tiwari[1] · **Bharti Nagpal**[2]

Received: 26 December 2021 / Accepted: 9 June 2022 / Published online: 11 July 2022

Abstract

Social media materialized as an influential platform that allows people to share their views on global and local issues. Sentiment analysis can handle these massive amounts of unstructured reviews and convert them into meaningful opinions. Undoubtedly, COVID-19 originated as the enormous challenge across the world that physically and financially bruted humankind. Meanwhile, farmers' protests shook up the world against three pieces of legislation passed by the Indian government. Hence, an artificial intelligence-based sentiment model is needed for suggesting the right direction toward outbreaks. Although Deep Neural Network (DNN) gained popularity in sentiment analysis applications, these still have a limitation of sequential training, high-dimension feature space, and equal feature importance distribution. In addition, inaccurate polarity scoring and utility-based topic modeling are other challenging aspects of sentiment analysis. It motivates us to propose a Knowledge-Enriched Attention-based Hybrid Transformer (KEAHT) model by enriching the explicit knowledge of Latent Dirichlet Allocation (LDA) topic modeling and lexicalized domain ontology. A pre-trained Bidirectional Encoder Representation from Transformer (BERT) is employed to train within a minimum training corpus. It provides the facility of attention mechanism and can solve complex text problems accurately. A comparative study with existing baselines and recent hybrid models affirms the credibility of the proposed KEAHT in the field of Natural Language Processing (NLP). This model emphasizes artificial intelligence's role in handling the situation of the global pandemic and democratic dispute in a country. Furthermore, two benchmark datasets, namely "COVID-19-Vaccine-Labelled-Tweets" and "Indian-Farmer-Protest-Labelled-Tweets", are also constructed to accommodate future researchers for outlining the essential facts associated with the outbreaks.

Keywords COVID-19 vaccine · Indian farmer protest · Bidirectional encoder representation from transformer (BERT) · Latent Dirichlet Allocation (LDA) · Lexicon approach · Social networks

✉ Dimple Tiwari
dimple.tiwari88@gmail.com

Extended author information available on the last page of the article

1 Introduction

The incremental growth in social networks changes the mechanism of interaction, communication, and information extraction. It plays a vital role in solving the serious issues people face worldwide. Indian farmer protest and the COVID-19 pandemic are two major challenging issues, where social networks played a crucial role in fighting against the situation. Even though the Twitter network is a major source of information and ideas exchange, these networks contain massive information. Therefore, sophisticated tools and techniques are required, which segregate the data and provide efficient decision-making [1]. Sentiment analysis is a popular approach to extracting meaningful emotion from unstructured tweets and reviews [2]. Coronavirus (COVID-19) became a global challenge in past years that has physically and economically shaken the world. Tens of millions of COVID-19 confirmed cases and more than thousands of death cases were reported worldwide by World Health Organization (WHO) [3]. Alongside the farmers' protest was the ongoing activity in the northern region of India against the three legislative bills passed by the Parliament in September 2020. A protest against rights is an integral part of a democratic country; social platforms play a leading role in conveying the demands and disaffection toward the government [4]. Unquestionably, the sentiment models of AI contribute to providing the right direction in these situations. People share different emotions and opinions through social media platforms during outbreaks. Therefore, computer scientists and researchers proposed several intelligent models for suggesting the right path for governments. The sentiment analysis task can be done by the lexicon approach [5], the supervised-learning approach [6], the unsupervised learning approach [7], and the deep learning approach [8]. All the techniques have their limitations in building an efficient sentiment analysis model. The increasing demand for social platforms for global outbreaks and the inefficiency of the existing sentiment models motivate us to build a hybrid model with lexicon-learning, unsupervised learning, and attention-based transformers. Existing sentiment methods train shallow models with carefully designed features and achieve satisfactory classification outcomes [9]. These models commonly applied classical algorithms, including Naïve Bayes (NB), Support Vector Machine (SVM), and clustering using linguistic attributes like Part-of-Speech (POS), Bag-of-Words (BoW), or linguistic features, which have two major drawbacks. First, the high-dimensional feature space minimizes the efficiency of the model. Second, the lengthy feature selection process takes more human effort and time [10]. Therefore, word embedding-based vectorization was introduced to overcome the drawbacks of traditional feature engineering techniques [11, 12]. The word embedding is a vector representation of the words that capture the word's semantic, syntactic, and contextual relationship with other words. These vectors feed into the DNNs as an input for further processing; it is increasingly favorable in recent NLP research. DNN models are mainly proposed to train machine learning tasks or a word embedding vectorization, including clustering and classification on calculated feature vectors. Over the past few years, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been gaining more admiration for text processing [13] due to the potential of CNNs

pattern learning and RNNs sequential modeling [14]. Despite the popularity of RNNs in text classification problems, they are experiencing the problem of exploding and vanishing gradients during the extensive text dependencies in a sentence. Such dependencies are usual in most of the reviews datasets and sentiment-related applications. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) were introduced to overcome the dependencies-related challenges. The LSTM and GRU are enriched by cell, input gate, output gate, forget gate, update gate, and reset gate. In this, the cell remembers the previous arbitrary values over time, and the three gates maintain the flow of information processing [15]. The GRU has an update and resets gate that decides which information should pass to the output with long-time storage information. They also remove the irrelevant information that does not require for the predictions. Due to their potential of relative insensitivity to large gaps, they gain attention in various NLP-based applications like speech recognition, named-entity recognition, handwriting recognition, and many more [16]. The Bidirectional GRU (Bi-GRU) and Bidirectional LSTM (Bi-LSTM) were introduced to read the sequence of left and right directions together. These models better solve the sequential text processing problems by integrating forward and backward hidden layers [17, 18]. Although these deep models are widely adopted in NLP applications and effectively solve text problems, still these have some loopholes, where three major challenges are unavoidable. First, the bidirectional networks hold double cells to remember previous and successive information, increasing complexity and extra overhead in producing effective outcomes during training. Second, the high-dimensionality input space distracts the focus of models from relevant contextual knowledge available in a text and makes them difficult to optimize. Third, deep models require a vast amount of data to train a model, and it is challenging for most real-life applications to train a model on millions of labeled data points.

The challenges mentioned above yielded the need for a sophisticated sentiment analysis model. Therefore, we proposed a KEAHT model for social sentiment analysis. To address the issue of inaccurate aspect–opinion pairs, the overhead of double cells, and small training corpus, we enrich the BERT transfer model with the dictionary-based sentiment polarity and LDA-based topic modeling. Moreover, we consolidate external knowledge like sentiment network graphs, text-length distribution, word count, and higher-polarity tweets to boost the performance of the BERT transfer model. To check the credibility, the proposed KEAHT has been investigated on two challenging issues (COVID-19 Vaccine and Indian Farmer Protests)-related tweets that belong to different domains. Experimental results reveal that the proposed hybrid model obtains promising results for the domain-specific small training corpus. The proposed model has been developed to fight against the COVID-19 outbreak and provides the right direction for the Indian farmer protest. Demonstrations reveal the credibility of social platforms and AI to fight against real-life global outbreaks. The major contributions of this research are the following:

- To propose a novel aspect-level hybrid KEAHT sentiment analysis model that is capable to retrieve accurate sentiment opinions for real-life situations with the highest accuracy score.

- To enrich the attention-based BERT transformer with extra-linguistic knowledge of lexicon-domain ontology, LDA topic modeling, and pair words network graphs.
- To introduce the real-life datasets ("COVID-19-Vaccine-Labelled-Tweets" and "Indian-Farmer-Protest-Labelled-Tweets") with informative fields for future research.
- To outline the informative global opinions related to the COVID-19 vaccine and farmer protest against the three legislative bills passed by the Indian government by exploiting the potential of AI.
- Conducted an extensive comparative analysis with baseline and existing hybrid models to evince the credibility of the proposed KEAHT against other state-of-the-art models.

The remainder subsections of this work are categorized as follows: Sect. 2 presents the systematic literature work of traditional classification models and neural models. Section 3 presents the preliminaries related to the proposed model. Section 4 displays the proposed methodology of BERT enrichment. Section 5 presents the experimented results. Section 6 shows the comparative discussion of the proposed model with existing baselines and recent hybrid models. Finally, Sect. 7 concludes the research and offers future directions.

2 Literature Review

Sentiment analysis is the widespread application of NLP and has been applied by researchers for solving numerous real-life challenges using different approaches. This section presents the systematic literature review of the previous work done in this field according to the applied techniques.

2.1 Lexicon Approach

The lexicon approach is the earliest sentiment analysis paradigm that calculates the semantic intention of the sentence or a document from a semantic orientation of the lexicons. The lexicon approach holds the dictionary of words related to the specific language and generates the polarity based on a particular sentence, aspect, or whole document. It uses adverbs and adjectives to mine the semantic orientation of the content. The process of tagging subjective phrases with their associated semantic intention comprises two fundamental approaches: dictionary-based and corpus-based. Dictionary-based approach judges the sentiments based on terms available in the lexicons and uses the online dictionaries to tag these words. On the contrary, the corpus-based approach relies on co-occurrences of syntactic or statistic patterns embedded in text corpora and a predefine set of positive and negative seed words [19, 20]. Semantic-Oriented CALculator (SO-CAL) calculates the semantic orientation of the words from the applied dictionary and predicts the intensification and negation of the phrase. SO-CALs perform well on completely unseen and

cross-domain datasets [5]. English language corpora are widely available, but the Arabic corpora and lexicons have rarely been found, leading to the need to generate Arabic language corpora for sentiment analysis [21].

The SentiCircles is a lexicon-based approach that has been proposed for Twitter sentiment analysis; it offers a static and prior sentiment polarity to text regardless of its aspect [22]. A novel WKWSCI sentiment lexicon has been developed that provides 69% of accuracy for Amazon product reviews and news headline datasets. It beats the performance of previously existing General Inquirer, Multi-perspective Question Answering (MPQA), National Research Council Canada (NRC), Hu & Lio opinion lexicon, and SO-CAL version 1.11 lexicons [23]. Table 1 presents the summarized view of the existing lexicon-based research. The lexicon-based approach also effectively offers the scores and slang word detection facility to classify multiclass problems [24]. A lexicon-based and corpus-based approach can be combined to obtain the optimized classification results [25].

2.2 Machine Learning Approach

The opinion datasets might contain subjective or objective information. Where subjective information includes positive and negative reviews, objective information holds facts regarding the text. Machine learning helps the system learn the task without prior programmed training [26]. The traditional sentiment classification models generally rely on supervised, semi-supervised, or unsupervised learning [27, 28]. Different feature selection methods like GainRatioAttributeEval (GR), ClassifierAttributeEval (CA), InfoGainAttributeEval, OneRAttributeEval (OneR), and Principal Component Analysis (PCA) are used for dimensionality reduction that helps in training better machine learning models [29]. The large labeled data are required for training the supervised models. Therefore, unsupervised dependency parsing model has been proposed for online text processing [30]. Spectral-clustering has been used to make the two different clusters (positive and negative) [31]. Table 2 presents the summary of previously existing machine learning-based research.

A hierarchical unsupervised approach is presented that combines univariate and multivariate distribution for anomalous online reviews detection [32]. The semi-supervised approach combines the features of the lexicon-based system

Table 1 Summary of existing lexicon-based research for sentiment analysis

Research	Dictionary approach	Corpus approach	Dataset type	Number of datasets
(Taboada et al. 2011) [5]	✓		Reviews	4
(Abdulla et al. 2014) [21]		✓	Tweets & Comments	2
(Kundi et al. 2014) [24]	✓		Tweets	1
(Saif et al. 2016) [22]		✓	Tweets	2
(Keshavarz et al. 2017) [25]	✓	✓	Twitter	6
(Khoo et al. 2018) [23]		✓	Reviews	2

Table 2 Summary of existing machine learning-based research for sentiment analysis

Research	Supervised	Unsupervised	Semi-supervised	Adopted Method	Dataset type	Number of datasets
(Fernández-Gavilanes et al. 2016) [30]		✓		Propagation	Reviews and Tweets	3
(Unnisa et al. 2016) [31]		✓		Spectral Clustering	Tweets	1
(Khan[2] et al. 2016) [34]			✓	SentiWordNet and SVM	Reviews	7
(Khan[1] et al. 2017) [33]			✓	Information Gain and Cosine Similarity	Reviews	1
(Kumar[2] et al. 2019) [32]		✓		Dirichlet-Multinomial Distribution	Reviews	1
(Rintyarna et al. 2019) [29]	✓			NB Multinomial, Bayesian Network, Multilayer Perceptron, J48, Logistic, Random Tree, and Random Forest	Reviews	2
(Kumar[1] et al. 2021) [26]	✓			Collaborative Filtering	Feedback	1

with machine learning-based techniques to optimize sentiment classification performance [33]. The incorporation of machine learning into the lexicon approach increases the domain independence in lexicons [34].

2.3 DNN and Transfer Learning Approach

DNN introduced better features than traditional supervised learners. They provide automated generation of the features, self-learning abilities, advanced analytics, and higher scalability. One major advantage of DNN over conventional models is that they highly depend on automated feature engineering. In contrast, machine learning algorithms require manual feature selection, which is time-consuming and computes incomplete features [35]. A combined CNN-LSTM model has been proposed with max-pooling, dropout, and batch processing features to acquire better sentiment analysis results [36]. LSTM and its variant GRU gain more attraction for text-based sentiment analysis among several DNN network models due to their long-length processing capability [10]. The CNN model can extract higher-level features with their max-pooling convolutional layers, where LSTM easily captures the long-distance dependency. Therefore, CNN-LSTM-based hybrid model has been proposed to increase the accuracy of movie reviews [37]. Table 3 summarizes existing research for sentiment analysis using deep learning and transfer learning approaches.

Although deep models offer several advantages over classical machine learning models, they still have the problem of multiple layers that require massive time for processing. The researchers introduced transfer learning to provide the solution to these problems. Transfer models are the pre-trained models that are further applied to different but related issues and can easily handle comparatively fewer data. Due to their pre-training, these can run fast and produce more accurate results than deep learning techniques [38]. In this work, we use the BERT transformer, an attention model that calculates the contextual relationships between word vectors in a text. However, transformers include encoders for reading the text inputs and decoders used for predicting the task. In contrast, BERT only contains encoders for generating a language model [39].

Table 3 Summary of existing deep learning and transfer learning-based research for sentiment analysis

Research	CNN	RNN	LSTM	Transformer	Dataset type	Number of datasets
(Daval-Frerot et al. 2018) [38]				✓	Twitter	1
(Rehman et al. 2019) [37]	✓		✓		Reviews	1
(Rani et al. 2019) [35]	✓				Reviews	1
(Basiri et al. 2021) [10]	✓	✓			Reviews + Tweets	8
(Singh et al. 2021) [39]				✓	Twitter	1
(Jain et al. 2021) [36]	✓		✓		Tweets	1

2.4 Background Limitations

The research related to opinion mining and sentiment analysis has been going on since the 2000s, and various paradigms have been introduced in the field of NLP to tackle the challenges related to text mining and opinion mining. However, the NLP advances in recent years with the latest AI techniques. Still, the literature has few pitfalls for critically analyzing the huge text available on social platforms. The major shortcomings of the background work for sentiment analysis are as follows:

- Most existing sentiment analysis models exploited machine learning and deep learning techniques to achieve higher text classification accuracy. The major challenge with the supervised algorithms is that they need a large corpus to train the model and provide slow performance.
- To overcome the problem of supervised classification techniques, unsupervised lexicon-based schemes are adopted by researchers for text sentiment classification. These methods are scalable, simple, and fast to process on large text, but they heavily rely on domain lexicons and are unable to acquire reliable results on cross-domain applications.
- Traditional sentiment approaches shallow train different techniques like SVM, NB, PoS, BoW, or clustering with carefully designed features and achieves satisfactory classification outcomes, which have two major drawbacks. First, the high-dimensional feature space minimizes the efficiency of the model. Second, the lengthy feature selection process takes more human effort and time.
- CNN and RNN have gained vast popularity for text processing due to their pattern learning and long-distance dependency over the past few years. Despite their wide adoption for NLP-related tasks, they have a vanishing and exploding gradient problem due to high dependencies between words. For solving these issues, GRU and bidirectional LSTM have been introduced, but they have the overhead of additional memory cells and increase the training complexity using high-dimensionality input space.

To overcome the shortcomings of existing sentiment analysis work, here, we proposed a novel model that can deal with the issue of inaccurate aspect–opinion pairs, the overhead of double cells, and a small training corpus. The proposed KEAHT is a hybrid sentiment model with the potential of supervised and unsupervised approaches.

3 Preliminaries

This section presents a brief overview of elementary components employed in the proposed KEAHT model. This model combines the potential of linguistic dictionaries, LDA-based aspect extraction, and attention-based BERT transformer.

3.1 Lexicon Dictionary

The lexicon-based method holds the dictionary of prior commonly used words. The sentiment score assigns to each word, and these words are then categorized into negative, positive, or neutral sentiments. After that, the manually collected words are searched into the lexicon dictionary based on the synonyms and antonyms. The collections of phrases, sentiment terms, or even idioms are also used to calculate the sentiment lexicons.

Figure 1 shows the procedure of lexical sentiment classification. In this work, we employed the TextBlob Application Programming Interface (API) for analyzing the sentiments of the COVID-19 vaccine and Indian farmer protest-related tweets. It facilitates NLP-related tasks, such as POS tagging, noun phrase extraction, sentiment analysis, and classification translation. The TextBlob extracts the sentiment from text based on two properties (Polarity and subjectivity); polarity means how much emotion is related to the positive (+1), negative (-1), and neutral (0) polarity. Where subjectivity identifies whether a sentence has emotion or not, it categorizes the text into two parts; if a sentence has fact, then it is declared as objective (0), and if a sentence has emotion, it is reported as subjective (1) [40]. Lexicon-based sentiment labeling is less sensitive to the data's quality and quantity, generating more accurate sentiment labeling [41].

$$\begin{aligned} &P(pos|w) \textit{for positive w} \\ &\text{P(neg | w) for negative w} \end{aligned} \tag{1}$$

$$\begin{aligned} P(pos|w) &= \frac{P(pos \cap w)}{p(w)} = \frac{\#wp}{\#w} \\ \text{P(neg | w)} &= \frac{P(neg \cap w)}{p(w)} = \frac{\#wn}{\#w} \end{aligned} \tag{2}$$

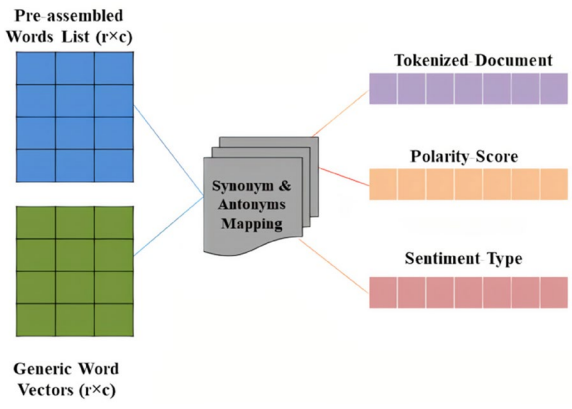


Fig. 1 Lexicon-based polarity calculation

$$F_N(s) = \left\{ \begin{array}{l} \max\left\{\frac{S+100}{2}, 10\right\} if\ \mathrm{s} < 0 \\ \min\left\{\frac{S-100}{2}, -10\right\} \text{ if } \mathrm{s} > 0 \end{array} \right\} \tag{3}$$

[Eq. 1] presents the conditional probability of each lexicon word. Conditional probability is calculated depending on the word negative or positive, shown in [Eq. 2]. Here, *#wp* indicates the number of positive sample words, and *#wn* presents the number of negative sample words. The most usual approach to negation handling in the text is reversing the polarity of the lexicon present just next to the negator, as presented in [Eq. 3] [42]. Here, F_N denotes the final negation, and S depicts the sentiment score from the adopted lexicon.

3.2 Aspect Extraction

Aspect extraction is a method of extracting the context from the views shared by the person. Topic modeling is a popular approach that recognizes the text's multiple topics and corresponding words. LDA is the most popular unsupervised, generative probabilistic model to extract highly recommended topics from a given corpus [43]. It works on two premises; first, the documents combine several topics, and second, the topics are the combination of different words. LDA has two parameters, α and β, where α manages the topic distribution for each document and β manages per topic word distribution.

Figure 2 describes the structure of LDA topic modeling. The affirmation algorithm in Table 4 depicts the process of LDA topic modeling for text documents [44].

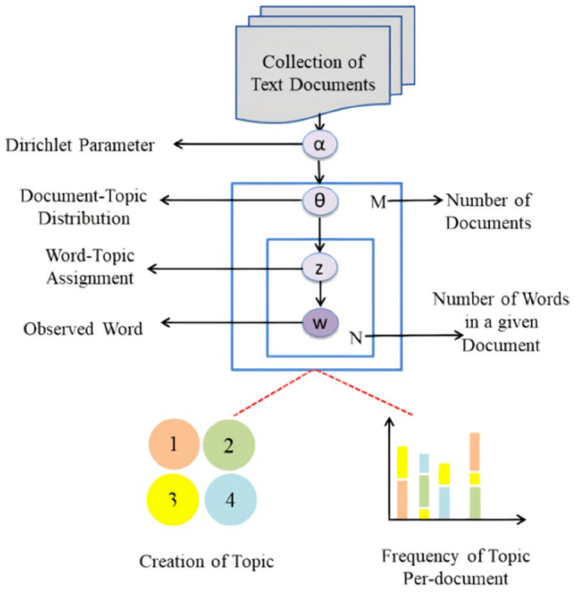


Fig. 2 The procedure of LDA topic modeling

Table 4 The LDA-based topic modeling calculation

Algorithm 1: An algorithm to perform LDA-based topic modeling

Input	Description
Corpus:	Dataset- $DS = \{ds_1, ds_2, \ldots .., ds_X\}$ #Where ds_X represents the document and X denotes the number of documents in the corpus
Document:	$W_x = \{w_{x,1}, w_{x,2}, \ldots, w_{x,Nx}\}$#Where, Nx represents the number of words in document d_x
Dictionary:	$.V = \{v_1, v_2, \ldots, v_V\}$. #Where V denotes the size of the dictionary
Latent:	$.Z_x = \{z_{x,1}, z_{x,2}, \ldots, z_{x,Nx}\}$. #Here, Z_x denotes the topic sequence associated with the word sequence W_x
	Process: 1. For all the topics $k \in 1, \ldots, K$: Choose a word distribution $\phi k \sim Dir(\phi\|\beta)$ 2. For all the documents d_x where x $\in$ 1, …, X •Choose $Nx \sim Poisson(\xi)$: •Choose a topic distribution $\theta m \sim Dir(\theta\|\alpha)$ •For all the words W_{Nx} where $n \in 1, \ldots, Nx$: Choose a topic index $zx, n \sim Mult(z\|\theta x)$ Choose a word $wx, n \sim Mult(w\|\phi zx, n)$

LDA effectively counts the words and groups according to similarity patterns within unstructured text.

$$p(\theta|\alpha) = \frac{r(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_{-1}} \ldots \theta_k^{\alpha_{k_{-1}}} \tag{4}$$

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) \tag{5}$$

A k-dimensional θ Dirichlet is a random variable that can assign values in (*k-1*) simplex if $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$ the probability density is calculated as in [Eq. 4], where the α is a k-vector of the components $\alpha > 0$ and $\Gamma(x)$ the Gamma function. Given the parameters α and β, the joint distribution of θ topic mixture, bunch of N topics z, and set of N words w is calculated in [Eq. 5]. Here $p(z_n|\theta)$ it simply shows the θ_i *for unique i like* $z_n^i = 1$ [45].

3.3 BERT Transformer

The BERT is an attention-based mechanism that takes deeper insight into the text and successfully finds the actual meaning of the content. It added the extra attention layer in previous DNN models to understand the context's importance better. BERT is the bidirectional trained model that reads the whole sentence in one call using the Masked LM feature; it uses the full context of the left and right directions to find the sentence's meaning. It is the advantage of the BERT model over existing DNNs, including CNN,

RNN, or LSTM. It is a pre-trained model that trains on a large text corpus and provides higher accuracy for different but related problems.

The BERT encoders take word vectors as inputs generated by the word embedding layer [46]. Figure 3 presents the additive attention mechanism, where the first step builds a junction representation of combined features calculated in timestamp t [47]. The embedded sentences [Eq. 6] and embedded aspect [Eq. 7] are as follows:

$$E = [e_1, e_2, ..., e_N]^T \in \mathfrak{R}^{N \times D} \tag{6}$$

$$E_A = [e_i, e_{i+1}, ..., e_{i+L}]^T \in \mathfrak{R}^{L \times D} \tag{7}$$

The sentence S further splits into two parts: S_{LS} starts from the sentence with the ending of the aspect [Eq. 8], and S_{RS} starts from the aspect with the end of the sentence [Eq. 9].

$$E_{LS} = [e_i, ...e_{i-1}, ..., e_{i+L}]^T \tag{8}$$

$$E_{RS} = [e_i, ...e_{i+L}, e_{i+L+1} ...., e_N]^T \tag{9}$$

After that, the attention mechanism guides the neural model for identifying the area, which requires more focus on a particular timestamp. Attention is a vector representation of a specified word/ token with esteemed all words/tokens in a queue. The attention output is named as context vectors, and context vector ct for timestamp t is calculated in [Eq. 10–13].

$$ht,n = \tanh(xt, Wa + xnWb) \tag{10}$$

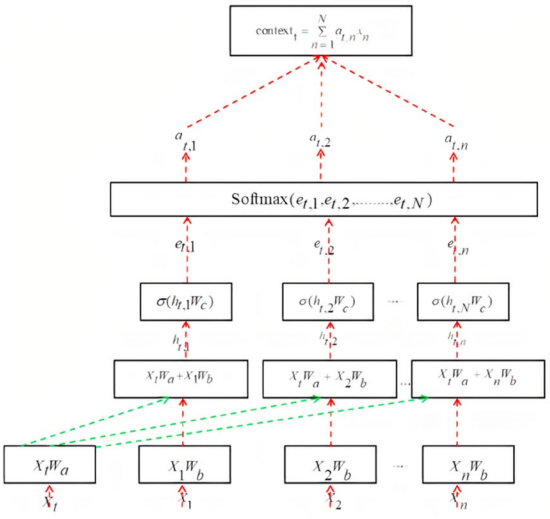


Fig. 3 An additive attention representation

$$e_{t,n} = \sigma(h_{t,n} W_c) \tag{11}$$

$$\mathrm{a_t} = \mathrm{softmax}(\mathrm{e_{t,1}}, \mathrm{e_{t,2}}, \ldots, \mathrm{e_{t,N}}) \tag{12}$$

$$\mathrm{c_t} = \sum_n a_{t,n} X_n \tag{13}$$

Here, for $n = 1$ to N, W is the weight vector, xt is a feature vector for timestamp t, and σ shows the sigmoid function. The BERT transformers remove the sequential nature of the previous DNN models and completely rely on the attention mechanism to calculate the dependencies between sequences. They use dot products and multi-head attention to achieve this purpose.

4 Proposed Methodology

To address the challenges of existing sentiment models. We developed a novel knowledge-enriched attention-based transformer model called KEAHT. Here first, we embedded the power of lexicalized word dictionary and the topic context modeling-LDA, then fed it into the BERT transformer for opinion extraction. Our research deals with two challenging affairs, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) vaccine and farmer protest against three legislated bills passed by the Indian government. The whole architecture of the proposed KEAHT model is presented in Fig. 4. The proposed KEAHT model holds a knowledge-enriched layer to extract the aspect and sentiment knowledge features using LDA topic modeling and lexicalized domain ontology. Given the input sentence $S = \{wd_0, wd_1, wd_2, \ldots\ldots\ldots, wd_n\}$ of length n, first, the extra domain-aspect knowledge is added as the input sentence $K = \{tp_0, tp_1, tp_2, \ldots\ldots\ldots, tp_n\}$ to guide the model's embedding. The obtained understanding of enriched sentence vectors can be represented as $I = \{v_0, v_1, v_2, \ldots\ldots\ldots, v_n\}$. Incorporating extra knowledge into the proposed KEAHT model helps in generating more reliable results. The basic BERT-based embedding layer acquires the raw sentences as input and trains on token-level representation, utilizing the information of the entire sentence. In contrast, the proposed KEAHT replaces the regular sentences with additional knowledge vectors in the form of utility topics. Similar to traditional BERT, the embedding layer learns an embedding orientation from additional knowledge vectors as the combination of token embedding, position embedding, and segment embedding. To feed into the BERT transformer, the input sequence is represented as $[CLS], v_0, v_1, v_2, \ldots\ldots\ldots, v_n, [SEP]$. Where [CLS] indicates the beginning of the sentence and the [SEP] token is used to separate the sentences from the subsequent one. Token embedding equipped the tokens with semantic and syntactic information. Then the transformer layer refines the token features layer by layer as $HI^n = Mask - Transformer_l(H^{n-1})$. Where the n mask-transformer layer has hidden size HI for knowledge input I. For the bert-base, there are 12 transformer layers, each of which contains 12 multi-head attention blocks. BERT uses 32 words of the

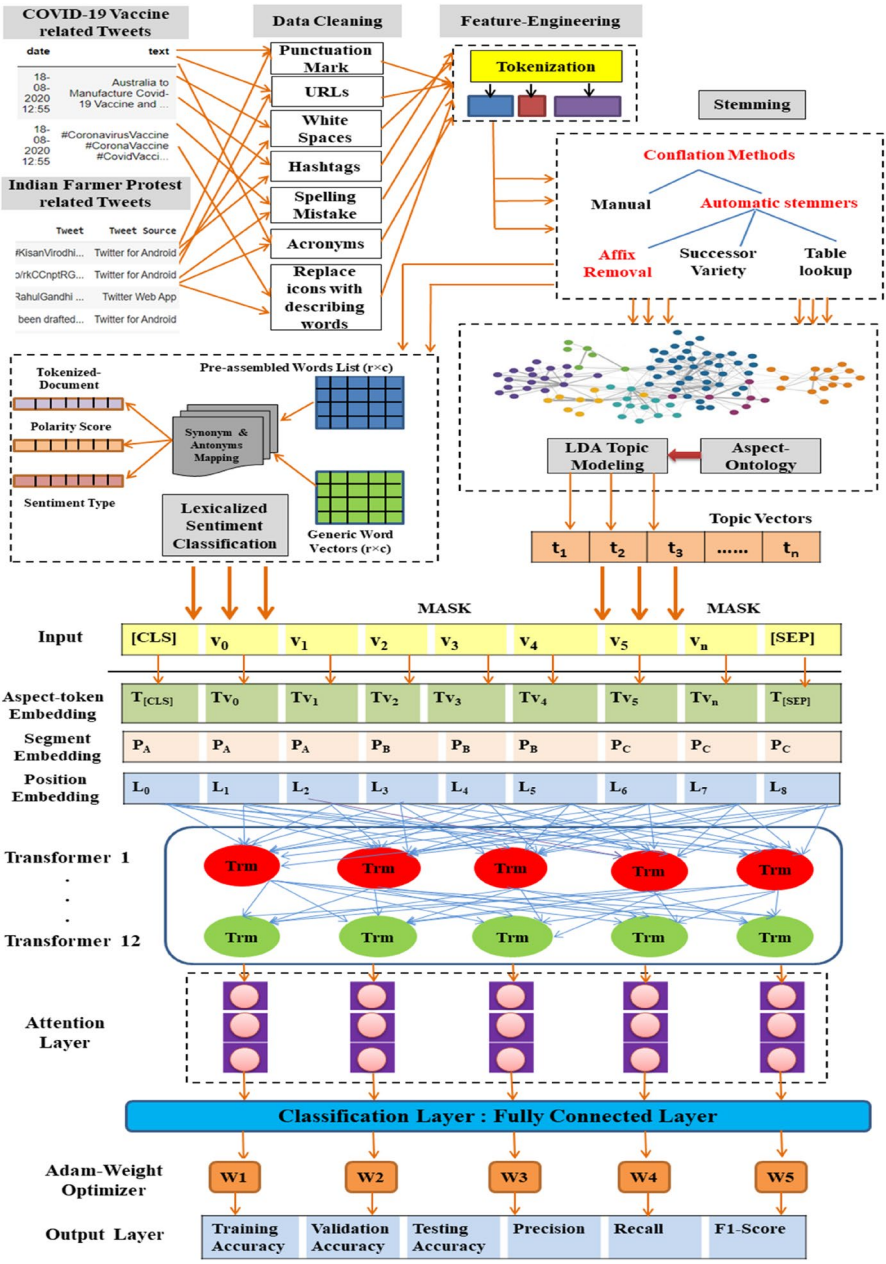


Fig. 4 The architecture of the proposed KEAHT sentiment model

input sentence as Masked Language Model (MLM) for future and previous context prediction. It obtains HI^n embedding orientation, passed through the linear attention layer to calculate the contribution of HI^n. After that, the fully connected layer

employs the softmax activation function to predict the sentiment polarity labels. Finally, the Adam weight optimizer is opted to correct the weight decay and improve the performance of the proposed model in terms of validation accuracy, training accuracy, testing accuracy, recall, precision, and f1-score. Table 5 presents the proposed hybrid algorithm with detailed formulated steps for sentiment analysis task. Comparative analysis with existing models confirms the credibility of the proposed KEAHT model. The further subsections present a brief overview of the conducted research.

4.1 Data Acquisition

The COVID-19 vaccine and Indian farmer protests-related tweets have been collected from the Kaggle (https://www.kaggle.com/datasets) and the GitHub (https://github.com/collections/open-data) data repositories. These two issues shook the people in the past years. Therefore, a need for the AI-based model has been arisen, which can spread awareness regarding the critical situations during the outbreak. Table 6 presents the information of the datasets before and after the knowledge enrichment. The COVID-19 vaccine dataset and Indian farmer protest dataset have 2987 and 3343 tweets, respectively, posted by different countries worldwide.

4.2 Pre-Processing

Data pre-processing is an essential step before starting the training process. It makes data ready to feed into a model for further processing [48]. Here, we perform the following steps to clean and normalize the datasets.

- **Noise Removal:** This step contains the basic noise-cleaning as whitespace, punctuations, URLs, and hashtag removal. Correction of spelling mistakes and acronyms has been done. After that, the emotion symbols have been replaced by their describing words [49].
- **Tokenization** is a fundamental step of NLP-related tasks in traditional count-vectorizer and advanced deep learning architectures. Here, the text or phrases are partitioned into smaller units called tokens. These smaller units can be either characters, words, or sub-word.
- **Porter-Stemmer:** is a procedure of separating common inflexional and morphological endings from the words. It extracts the similar or stemmed words that are semantically similar and then replaces them with single root words [50].

The pre-processing techniques enhance the quality of text datasets and convert them into machine-readable form.

Table 5 The proposed KEAHT algorithm for sentiment analysis

Algorithm 2: The proposed KEAHT algorithm for social sentiment analysis

Step1: Collect COVID-19 vaccine and Indian farmer protest-associated raw tweets

Step2: Examine each point of data and perform basic noise removal, tokenization, and stemming process as follows:

Tokenization: After completing the basic noise removal and normalization of the content. First, the sentence is broken into small chunks called tokens. For generating the tokens, the probability of a sentence is calculated by multiplying the probability of sub-words

Tokenization-$P(x) = \prod_{i=1}^{M} p(x_i)$,

Here, x is a sentence, *and* x_i is a Sub-word forming sentence. It is assuming that the number of words ranges from x_i,……,x_M

Vocabulary Generation: By applying tokenization, sentences are broken into smaller units called tokens, and these tokens are used to create a word vocabulary from the training corpus

$\forall_i x_i \in v, \sum_{x \in v}^{n} p(x) = 1$# V: vocabulary

Note: Heuristically, a giant seed dictionary must generate from the training corpus

Stemming: Once the tokenization is completed, the porter-stemmer method is applied to reduce the extra words from the vocabulary by removing affixes and suffixes

Porter-stemmer- (m > 1 and (*s or *T))

Where test for a stem with m > 1 ending in S (stem) or T

Step 3: Aspect extraction is very much required to extract the context from the views shared by the people. For this, LDA topic modeling is implemented, which is the most popular unsupervised, generative probabilistic model to extract highly recommended topics from a given corpus

$$\text{LDA} = (Z_i = j | z_{-i}, w, d_i, .) \propto \frac{C_{w_{ij}}^{WT} + \beta}{\sum_{w=1}^{w} C_{w_{ij}}^{WT} + W\beta} \frac{C_{d,j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d,t}^{DT} + T\alpha} \quad \text{\# Topic 1 probability}$$

Where $(Z_i = j | z_{-i}, w, d_i, .)$ is the probability of the topic for a word in a document, $C_{w_{ij}}^{WT} + \beta$ calculates the probability of word w in topic t, $\sum_{t=1}^{T} C_{d,t}^{DT} + T\alpha$ and calculates the probability of topic t in document d. Here, β represents the topic distribution per topic, α is a per-document topic distribution, W is a vocabulary length, T represents the number of topics, and $C_{d,j}^{DT}$ shows the iteration (counted the occurrence of the document as topic 1 and topic 2)

Step 4: Then unsupervised lexicon dictionary method is employed to extract the sentiment polarity from the knowledge-enriched dataset

$P(pos|w)$*for positive w*

$P(neu \mid w)$*for neutral w*

$P(neg|w)$*for negative w*

The conditional probability of each word lexicon in the sentence is calculated

Step 5: The BERT word embedding is performed to create the word vectors of the knowledge-enriched dataset. Hence, the embedded aspect and sentence has given as:

1. $E_A = [e_i, e_{i+1}, \ldots\ldots\ldots\ldots e_{i+L}]^T \in \mathbb{R}^{N \times D}$

$E = [e_1, e_2, \ldots\ldots\ldots\ldots e_N]^T \in \mathbb{R}^{N \times D}$

Proportionately, sentence S splits into two sections, *S[LS]* and *S[RS]*. Their corresponding representation is as:

$E[LS] = [e_1, \ldots\ldots\ldots e_{i-1}, e_i \ldots\ldots\ldots\ldots, e_{i+L}]^T$

$E[RS] = [e_1, \ldots\ldots\ldots e_{i+L}, e_{i+L+1} \ldots\ldots\ldots\ldots, e_N]^T$

Here, *S[LS]* starts from the sentence with the ending of the aspect, and *S[RS]* starts from the aspect with the ending of the sentence

Step 6: Bert bidirectional transformer is applied to find the relationship among all words in a sentence regardless of their position

Step 7: An attention layer has been connected to the network to generate weighted vectors for identifying the current vector that requires more attention on a given timestamp. The layer's head calculates α as attention weight between entire pair phrases like softmax-normalized dot products

$\alpha_{ij} = \frac{\exp(q_i^T k_j)}{\sum_{t=1}^{n} \exp(q_i^T k_t)}$ # attention weights

$o_i = \sum_{j=1} \alpha_{ij} v_j$ # the attention head output o

Step 8: Then, BERT-based sequence classification has applied to categorize the sentence as *pos* (positive), *neg* (negative), or *neu* (neutral)

Step 9: The Adaptive Moment Estimation (Adam) weight decay optimizer has been applied to improve classification accuracy. The decaying averages of previous and previous squared gradients (m_t and v_t) are calculated as follows:

$m_t = \beta 1 m_{t-1} + (1 = \beta 1) g_t$

$v_t = \beta 2 v_{t-1} + (1 - \beta 2) g_t^2$

Here, m_t and v_t calculate the first and second mean of the gradients, respectively

Step 10: Finally, the results have been declared in terms of standard measures, including validation accuracy, testing accuracy, training accuracy, model loss, recall, precision, f1-score, and corresponding confusion matrix

Table 6 The attributes of the tweets datasets before and after the knowledge enrichment process

Dataset	Domain	DataType
COVID-19 Vaccine	User_Name	object
	User_Location	object
	Date	object
	Text	object
	Hashtags	object
Indian Farmer Protest	Date	object
	Tweets	object
	Source	object
	Location	object
Domains added after the knowledge enrichment		
COVID-19 Vaccine & Indian Farmer Protest	Clean-Tweet	object
	Text_Length	int64
	Word_Count	int64
	Polarity	float64
	Sentiment_Type	object

4.3 Data Visualization

This section presents the extra information on the COVID-19 vaccine and Indian farmer protest-associated tweets. A deep analysis of people's attitudes toward these challenging issues is required. Therefore, we highlighted the crucial facts that spread the foremost truth about the circumstances. Table 7 depicts the highly polarized positive tweets associated with the COVID-19 vaccine, Table 8 presents the highly negative tweets associated with COVID-19 vaccine, Table 9 displays the highly polarized positive tweets associated with the Indian farmer protests, and Table 10 presents the highly negative polarized tweets associated with Indian farmers protests.

Figure 5a, b presents the sentiment polarity distribution of the COVID-19 vaccine and Indian farmer protest-related tweets. It has clearly shown that most of the tweets are classified as neutral for both situations, which conveys that people have a neutral opinion toward the COVID-19 vaccine and for the Indian farmer protests.

Figure 6a presents the cumulative word count of the COVID-19 vaccine-related tweets. Most people have posted approx. 15 to 20 words to show their opinion. Figure 6b illustrates the cumulative word count for Indian farmer protest-related tweets, and here most people have posted around ten words to share their views.

All the words initiated with the symbol hash (#) are called hashtags. These tags are useful for recognizing trending affairs. Figure 7a, b depicts the word cloud of common hashtags available in the COVID-19 vaccine and Indian farmer protests associated tweets. The word clouds represent the image of text, where the size of the phrase is proportionate to the prevalence of occurrence. Therefore, the more prominent phrase expresses the most frequent words posted in tweets [51]. At a glance, the word cloud of COVID-19 vaccine shows more common words discussed by

people such as "Covid-vaccine", "Russia", "First Covid", "people", and "coronavirus", "trial", and "health ministry". In the case of Indian farmer protests, excessive tweeted hashtags are "farmer", "support", "protest", "India"," Modi"," bill", "government", "law"," leader" and "country". The most frequent tweets constitute the people's actual thought process for a specific issue.

4.4 Aspect Extraction with Topic Modeling

Topic modeling is the most recommended technique in NLP for text mining, latent subject discovery, and pair words relations finding [52]. Aspect extraction is a method of extracting the context from the views shared by the person [53]. Although various techniques are available to extract more relevant topics from the content, LDA is one of the most robust methods in this field. Table 11 presents the algorithm of ontology-based LDA topic modeling.

LDA is an unsupervised generative probabilistic approach; it assumes that each document can be drafted as a probabilistic distribution over latent topics. In all documents, topic distribution allocates a common Dirichlet prior. This provides the computational content analysis to investigate the "unseen" thematic construction of the collected text [54]. Therefore, we applied this to identify the major aspects discussed by people for the situations. Figures 8 and 9 depict the dependency between topics and words of COVID-19 vaccine and Indian farmer protests associated tweets. The above network graph shows the relationship between the extracted topics and

Table 7 Two random tweets from the COVID-19 vaccine dataset with the highly positive sentiment polarity

Tweet Number	Highly Positive Tweets
1	*"We have the best research labs"*
2	*"Excellent news! It shows how diligent we've all been."*

Table 8 Two random tweets from the COVID-19 vaccine dataset with the highly negative sentiment polarity

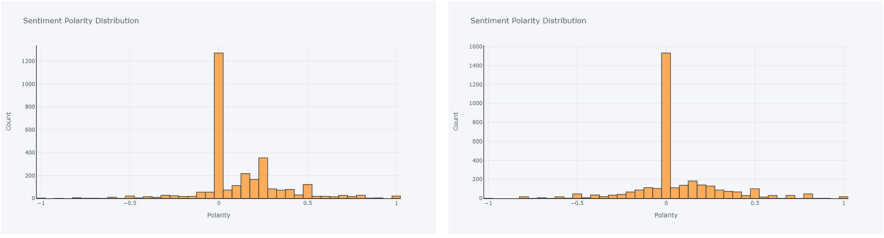
Tweet Number	Highly Negative Tweets
1	*"India's COVID tally crosses 29L-mark, toll nears 55"*
2	*"Russia Coronavirus Vaccine Not Sufficiently Tested? Germany Questions Safety of World"*

Table 9 Two random tweets from the Indian farmer protests dataset with the highly positive sentiment polarity

Tweet Number	Highly Positive Tweets
1	*"its the best analysis they have read date"*
2	*"farmers are not anti-national they care for the country which why they are protesting for their fundamental rights procuring the best prices for their crops"*

Table 10 Two random tweets from the Indian farmer protests dataset with the highly negative sentiment polarity

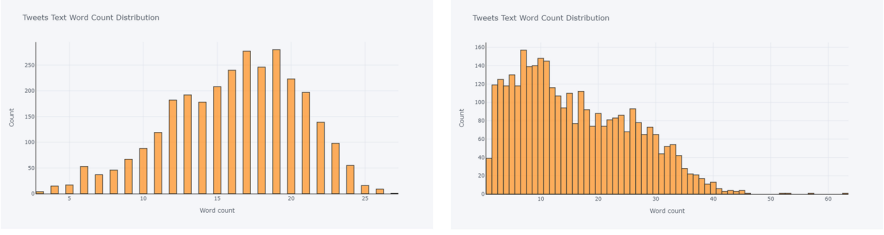
Tweet Number	Highly Negative Tweets
1	*"the way you oppose Khalistan but respect Sikhism same way oppose Hindutva but respect Hinduism sums the ga-ndhi-vadra family drama"*
2	*"once your life you need doctor lawyer a policeman a preacher but every day three times day you need farmer how the government calculating minimum support price msp kaneesa maddatu dhara the trick lies there please explain this consensus out issues reached between farmer"*



(a) Sentiment polarity distribution of Covid-19 vaccine-associated tweets

(b) Sentiment polarity distribution of Indian farmer protests associated tweets

Fig. 5 Sentiment polarity distribution of collected tweets

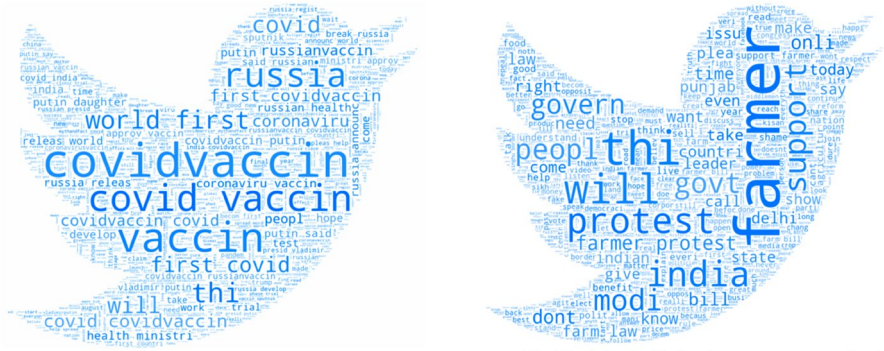


(a) Word count distribution of Covid-19 vaccine-associated tweets

(b) Word count distribution of Indian farmer protests associated tweets

Fig. 6 Word count distribution of collected tweets

the frequency of their related words. Twitter connects people of different countries who share opinions from different locations, leading to the semantic heterogeneity problem in divergent environments. Therefore, we used the network graph to find the correspondence among the views of entities belonging to different regions [55].



(a) Word cloud of COVID-19 vaccine-associated tweets

(b) Word-cloud of Indian farmer protests associated tweets

Fig. 7 Word cloud calculated for the tweets

Moreover, it is clear that the people of different countries have shared almost the same opinions and used the same hashtags (context) toward both severe issues. People shared neutral views on the COVID-19 vaccine and Indian farmer protest-related issues.

4.5 Lexicalized Sentiment Extraction

A TextBlob API has been used to generate the subjectivity and sentiment polarity of the raw tweets. It extracts the exact sentiment class of the tweets based on the sentiment polarity score. The polarity score is a float value, which lies within [-1.0, 0.0, 1.0]. Here, -1.0 indicates the negative, 0.0 indicates the neutral, and 1.0 represents the positive sentiment class. The subjectivity score also produces a float value that lies within the range of [1.0, 0.0], where 1.0 indicates very subjective, which means a text holds an emotion, and 0.0 indicates very objective, which means the text holds a fact. These polarity scores and the calculated sentiments were then fed into the attention model to calculate the classification score in terms of accuracy, recall, precision, and f1-score. Table 12 presents the process of lexicalized dictionary-based sentiment calculation.

Our primary objective of the research is to enrich the attention-based transformer model with combined features of aspect ontology and lexicalized dictionaries. Therefore, we perform the aspect extraction using the LDA modeling technique and present the ontology-based network graphs for employing the potential of an unsupervised approach. We have also calculated the polarity score of the tweets for sentiment extraction.

4.6 BERT Training

After completing the knowledge enrichment process, we trained the BERT transformer on our COVID-19 vaccine and Indian farmer protest-associated datasets.

Table 11 LDA topic modeling with network-based ontology graph

Data: Corpus-1 (CV): COVID-19 vaccine-associated tweets
Corpus-2 (IFP): Indian farmer protest-associated tweets
Results: Topic with associated word-features
Begin:
1.# LDA-based topic modeling
2.Consider N sequence of words in document D
3.**for each** topic ti ϵ {1,2,3,4,…..TI} **do**
4.| Select a word distribution $\Phi_{tp} \sim Dirichlet(\beta)$;
5.**end for**
6.**for each** doc ϵ corpus CV & IFP **do**
7.| Consider the topic distribution $\Phi_d \sim Dirichlet(\alpha)$;
8.| **for** each N word of doc (W_{di}) **do**
9.| | sample a topic $TI_d \sim Multinomial(\theta_d)$;
10.| | sample a word W_{di} from $p(w_n|TI_d)$;
11.| **end for**
12.**end for**
13.# Network ontology representation of Topic related words
14.**for each** topic ti
15.| **for each** word W_{di}
16.| | **if** Wdi in word of TI
17.| | | TI = TI + 1;
18.| | **end if**
19.| **end for**
20.**end for**

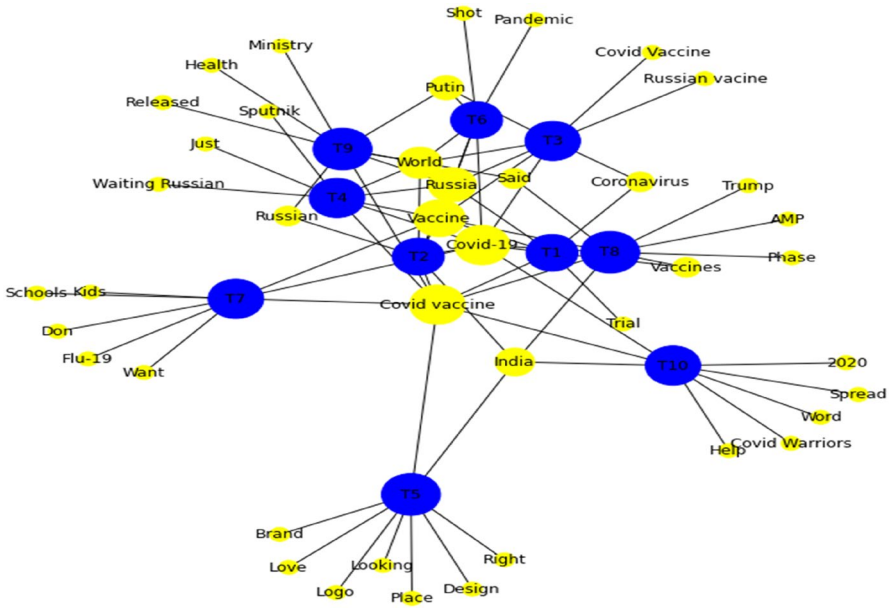


Fig. 8 Network diagram of topic and word extracted from COVID-19 vaccine-associated tweets

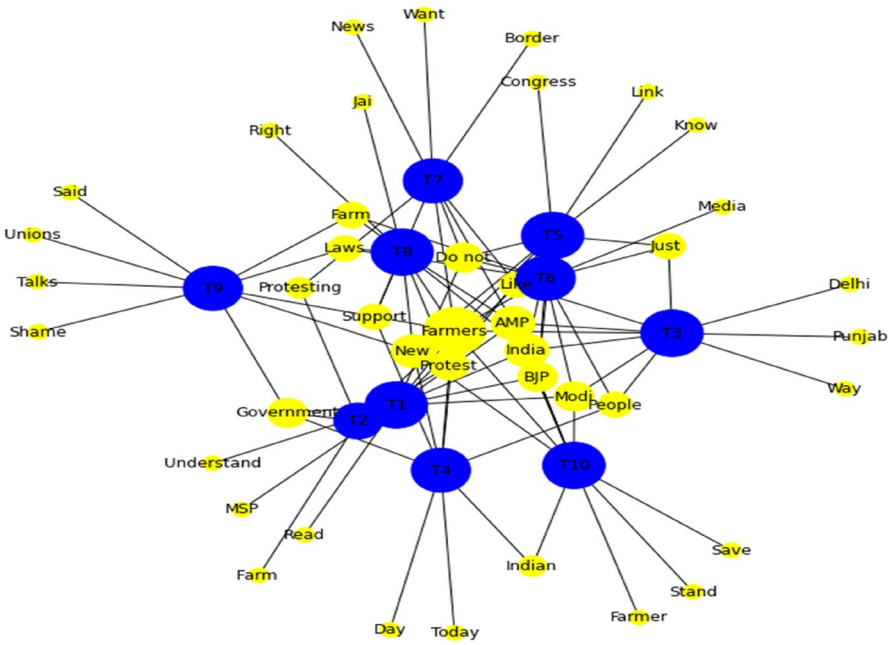


Fig. 9 Network diagram of topic and word extracted from Indian farmer protests associated tweets

These datasets now hold extra domain values due to the knowledge enrichment process shown in previous sections. Moreover, this supplementary knowledge has fed into the BERT transformer, which helps in increasing the classification accuracy and produces reliable results. To facilitate our model, we installed the transformer 3.0.2, CPython 3.7.6, IPython 7.13.0, NumPy 1.18.5, pandas 1.1.2, and torch 1.0.6. We applied the bert-base-uncased model for the experiments; it holds 12 main layers, 768 hidden layers, 12 heads, and 110 M parameters for NLP-based training. The detailed experimental steps of BERT training are as follows:

- A BertTokenizerFast has been used to implement the word embedding. Here, we added the unique tokens, namely [SEP] for ending marker, [CLS] for special token starting identification, [PAD] for sentence specific length padding marker, and [UNK] for a token marker, except tokens in trained data. We have chosen the maximum length of 32 for the attention mask. Figure 10a, b presents the token count for the COVID-19 vaccine and Indian farmer protests-related tweets.

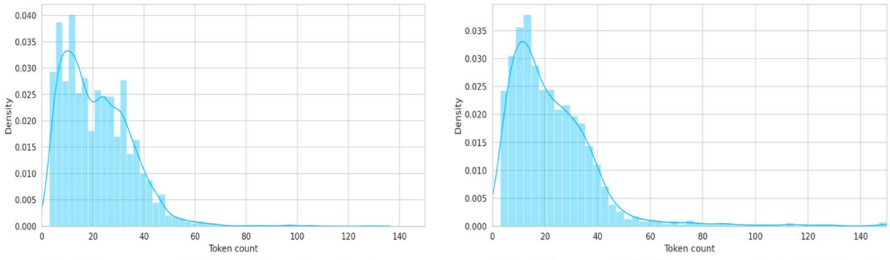
  The presented graphs show that the maximum length of most tweets lies around 6 to 32 words in both the datasets, but some go around 150. Therefore, we have chosen a full sequence length of 150 for both datasets.

Table 12 Sentiment extraction with lexicalized dictionary approach

Algorithm 3: An algorithm to extract sentiment from tweets by lexicalized dictionary approach

Data: Generic Word Vectors (GW)—COVID-19 vaccine tweets or Indian farmer protest tweets
Preliminaries: Pre-assembled Word List (WL)- TextBlob
Begin:
1.analysis = WL(GW)
2.**For each** row in analysis
3.| **if** analysis.Polarity-Score > 0
4.| | sentiment = "Positive";
5.| **elif** analysis.Polarity-Score < 0
6.| | sentiment = "Negative";
7.| **elif** analysis.Polarity-Score = 0
8.| | sentiment = "Neutral"
9.| **end if**
10.**end for**
Output: Sentiment- Negative, Neutral, or Positive

- After the word embedding and masking, we split the dataset into 80% for training and 20% for testing purposes. The testing part is further broken into a 50:50 ratio; 50% for validation and 50% for testing.
- Then the helper function has formulized for the data loader, which holds the data frame, tokenizer-length 32, sequence length 150, and batch size 32. The data loader merges the dataset with the sampler and dispenses an iterable set over the experimented dataset.
- After that, the classification layer has created using the pre-trained BERT model. For a summary of the content, BertPooler was applied to last_hidden_state. Then we connected the dropout layer to regulate the model complexity and the fully connected layer for output generation.
- Then the softmax function was created to predict the probabilities by the trained model. This was the final activation function for the classification problem. It converts the input vectors into probability vectors, where the possibilities of each value are proportional to the related scale of each value in the vector. The softmax activation is calculated in [Eq. 14].



(a) Token count of COVID-19 vaccine-related tweets (b) Token count of Indian farmer protest-related tweets

Fig. 10 Token counts of collected tweets

$$\sigma(\vec{S})_i = \frac{e^{s_i}}{\sum_{j=1}^{n} e^{s_i}} \tag{14}$$

Here $(\vec{S})$ is an input vector, *si* represents the value of the input vector, $e^{s}{}_{i}$ is an exponential function applied to each input vector, and *n* represents the number of classes used in multiple classifications [56]. It normalizes the input vectors into ranges that often generate probabilistic interpretations.

- The Adam weight optimizer has been applied to correct the weight decay that enhances the results of the proposed model. It computes the decaying averages of previous and previous squared gradients m_t and v_t as in [Eq. 15] and [Eq. 16], respectively.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{15}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{16}$$

Adam basically depends on two values β_1 and β_2. Where β_1 is the exponential decay of the rate for the first-moment estimates and β_2 is for second-moment estimates. Moving averages are calculated using β_1,β_2, and g_t parameters on given iteration t. Almost all the algorithms of moving averages are biased. Thus, the extra step has been added to correct the bias. Then, they counteract the above biases by calculating bias-corrected first and second-moment estimates in [Eq. 17] and [Eq. 18], respectively.

$$m_t = \frac{m_t}{1 - \beta_1^t} \tag{17}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{18}$$

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{\hat{v}_t} + \in} m_t \tag{19}$$

Finally, weights and biases are updated in [Equation 19] by moving averages calculation. It computes the adaptive learning rates for each parameter [57].

- Finally, the helper function has created to evaluate the model, which was initialized with three epochs for model learning. The model generated higher accuracy and lower model loss at the final epoch. Therefore, we stopped the learning at the third iteration.

5 Experimental Results

The motivation behind the proposed KEAHT model is to build an AI-based model, which can monitor the outbreak situation and provide the right direction to the nation in dispute. Therefore, the experiments were conducted on two major challenging issues (COVID-19 vaccine and Indian Farmer protests), which need intelligent suggestions.

5.1 Evaluation Criteria

Here, we use a few standard measures to check the performance of the KEAHT model. The confusion matrix calculated the results, also known as an error matrix. The confusion matrix is an N * N matrix, which compares the actual target values with the predicted values by the model. It generates a tabular view to visualize the performance of the algorithm.

Table 13 represents the structure of the confusion matrix. A target variable holds two values, namely positive and negative. In the confusion matrix, columns depict actual values, and rows represent the predicted values by the model. Here, TP (True Positive) means actual positive predictions, TN (True Negative) means actual negative predictions, FP (False Positive) means incorrect positive predictions, and FN (False Negative) means incorrect negative predictions. We have created the classification report, where we calculated the accuracy, precision, recall, and f1-score for the proposed model. The formulas of these measures are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{21}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{22}$$

$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{23}$$

Here, accuracy shows the ratio of correctly predicted outcomes to the total outcomes, precision is the ratio of correctly predicted outcomes to the total predicted

Table 13 Confusion matrix

Actual values			
Predicted values		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

positive outcomes, recall is the ratio of correctly predicted outcomes to the total actual positive outcomes, and f1-score is the harmonic mean of precision and recall [58].

5.2 Results and Discussion

The idea behind the proposed model is to integrate the potential of different intelligent approaches, including lexicalized dictionaries, unsupervised learning, aspect extraction, and attention mechanism, to fight against real-life pandemics and inconvenient situations. This section presents the results obtained by the proposed KEAHT model to deal with COVID-19 and Indian farmer protests. Moreover, the results achieved by the proposed solution confirm the validity of the model for upcoming outbreaks. Table 14 presents the positive, negative, and neutral sentiment count calculated by the lexicalized dictionary TextBlob.

It has visualized that the people share positive and neutral opinions toward both issues. Table 15 presents the associated aspect-topics calculated by the LDA modeling. The highlighted words show the uncommon words available in the topics, and the rest are very common; most people have common aspects toward the COVID-19 vaccine and Indian farmer protest.

Figure 11a, b presents the frequency of the most common words available in the COVID-19 vaccine and Indian farmer protest-associated tweets. It represents the major aspects that are commonly discussed over the world for selected real-time issues. After completing all the above lexicalized dictionaries and the unsupervised aspect extraction-related tasks, the datasets have been updated with extra-enriched knowledge. Consequently, the attention-based transformer has been trained on knowledge-enriched datasets to classify the nation's sentiments toward the state of affairs. Table 16 presents the accuracies and losses of the proposed KEAHT model. It has shown that the proposed model achieves 92.84% training and 90.63% testing accuracy for COVID-19 vaccine-related tweets, which stated the achievements of the knowledge-enriched solution. For Indian farmer protest-related tweets, our model scores 92.63% training and 81.49% testing accuracy. Furthermore, the validation phase model has been hyper-tuned effectively for both cases, and it produced almost similar validation and testing accuracy scores.

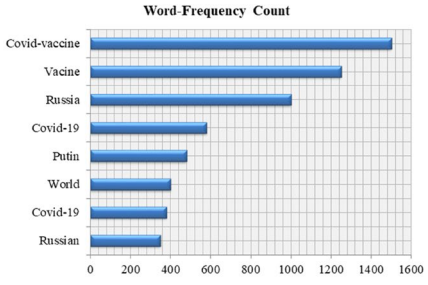
Figure 12a, b presents the training history of the proposed model during the three iterations. It shows that the model continuously enhanced its performance over the iterations, but it became stable at the last iteration. Here, we stopped training, started testing on the test set, and achieved higher accuracy for sentiment classification of the tweets, as presented in Table 16. Figure 12b, d illustrates the training loss of the model, which has continuously decreased over the iterations; it shows that the model's performance has continually improved, and failure has constantly reduced

Table 14 Sentiment count calculated by lexicalized dictionary

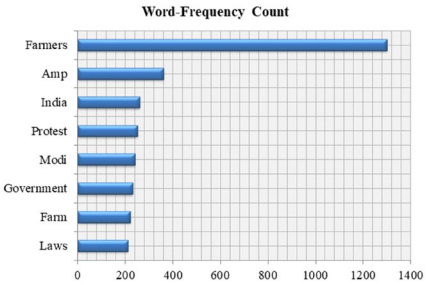
Dataset	Positive	Neutral	Negative	Total
COVID-19 Vaccine	1465	1249	274	2988
Indian farmer protest	1239	1483	621	3343

Table 15 Aspect associated topics calculated by the LDA topic modeling

COVID-19 Vaccine-associated tweets	
Topic-1	Covid-vaccine, Vaccine, COVID-19, Russia, Coronavirus, Vaccines, **Trial**
Topic-2	Covid-vaccine, Vaccine, COVID-19, Russia, World, India, Russian,
Topic-3	Russia, Vaccine, Covid-vaccine, Putin, COVID-19, Russian vaccine, World, Coronavirus,
Topic-4	Covid-vaccine, Russia, Vaccine, COVID-19, World, **Just**, **Waiting Russian**, **Sputnik**
Topic-5	Covid-vaccine, **Place**, **Brand**, **Design**, **Right**, **Love**, **Logo**, **Looking**, India
Topic-6	Vaccine, COVID-19, Russia, **Pandemic**, Putin, **Shot**, World
Topic-7	Covid-vaccine, vaccine, **Kids**, COVID-19, **Want**, **Don**, **Schools**, **Flu-19**
Topic-8	Covid-vaccine, Vaccines, COVID-19, India, **Trump**, Said, **Phase**
Topic-9	Covid-vaccine, Russia, Russian, **Health**, World, Said, Putin, **Ministry**, **Released**
Topic-10	**Help**, Covid-vaccine, COVID-19, India, **Word**, **Spread**, **Covid Warriors**
Indian farmer protests associated tweets	
Topic-1	Farmers, Modi, Government, AMP, **MSP**, **Like**, **Understand**, **Read**, Do Not, BJP
Topic-2	Farmers, Protest, India, Government, Farm, Protesting
Topic-3	Farmers, AMP, India, Modi, **Delhi**, People, **Way**, Just, Like, **Punjab**
Topic-4	Farmers, Support, Day, Indian, People, Protest, Government, **Today**, New
Topic-5	Farmers, Just, **Know**, India, Do Not, Protest, BJP, **Link**, **Congress**, New
Topic-6	Farmers, People, Modi, Farm, Laws, **Media**, Do Not, Just, BJP
Topic-7	Farmers, Like, Support, **Border**, Do Not, Protesting, News, **Want**, AMP,
Topic-8	Farmers, AMP, India, Laws, Farm, New, Support, **Right**, **Jai**, Protest,
Topic-9	Farmers, Laws, Farm, Talks, Government, **Unions**, **Shame**, New, **Said**
Topic-10	Farmers, AMP, India, BJP, **Stand**, Indian, Modi, New, **Save**



(a) Word-Frequency count of COVID-19 vaccine-associated tweets



(b) Word-frequency count of Indian farmer protest associated tweets

Fig. 11 Word-frequency count of collected tweets

till the last iteration. Finally, we obtain the highest version of the proposed model. Figure 13a, b depicts the confusion matrix of the proposed model on COVID-19 and Indian farmer protest datasets, respectively. Table 17 shows the precision, recall, and f1-score obtained by the proposed model. It is demonstrated that the model achieved

Table 16 The accuracy (Acc) and loss obtained by the proposed KEAHT model

Dataset	Training-Acc	Validation-Acc	Testing-Acc	Training Loss	Validation Loss
COVID-19 Vaccine	92.84%	89.29%	91%	23.68%	34.64%
Indian farmer protest	92.63%	81.43%	81.49%	23.08%	54.75%

higher classification results for COVID-19 vaccine-related tweets than Indian farmer protests.

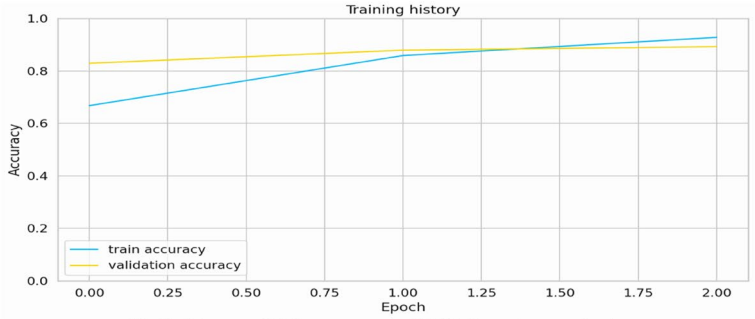
6 Comparative Analysis

A comparative analysis is required to confirm the credibility of the proposed model. This section presents a close, in-depth discussion of the proposed model with baseline models and existing approaches. We compared our hybrid model with existing machine ensembles and deep learning-based techniques. Furthermore, the current hybrid models have also been compared to prove the novelty and accomplishment of the proposed hybrid model.

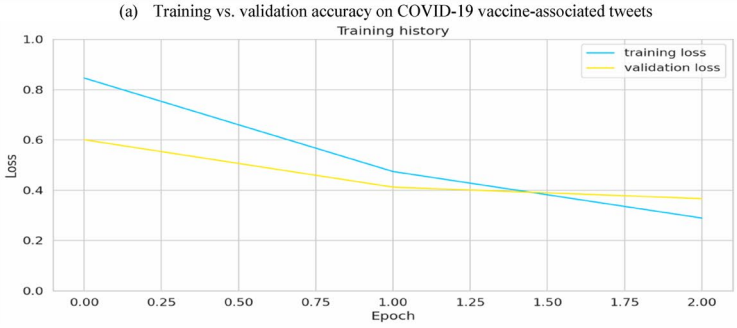
6.1 Comparison with Baseline Models

Sentiment analysis has been a popular application for past decades. Machine learning and deep learning continuously produce significant results for analyzing people's sentiments on different social platforms, but both have several limitations. Therefore, we built a hybrid attention-based model to handle the real-life situation and combine the potential of both the existing approaches. This subsection individually compared the proposed model with machine learning and deep learning-based techniques. Table 18 presents the comparative results of the proposed hybrid model with existing machine ensembles and deep learning-based methods. The results obtained by the proposed KEAHT model are highlighted in bold. We evaluated the influential machine ensembles (Ada Boost, Extra Tree, Gradient Boosting, Random Forest, Extreme-Gradient Boosting, and Light GBM) techniques on our datasets. Moreover, it has clearly shown that our proposed hybrid model outperformed both the real-time datasets and generated higher results than baselines in terms of accuracy, precision, recall, and f1-score.

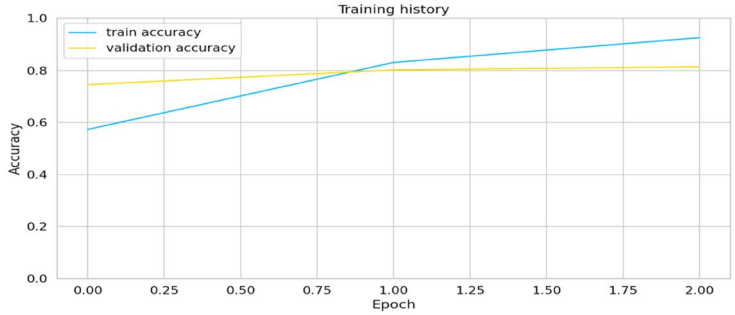
Furthermore, we have been evaluated popular neural models like Sequential RELU, Deep CNN, and RNN LSTM. Demonstration depicts the below performance of these models for both real-life circumstances. Although DNNs effectively solve complex text problems but require a vast amount of datasets for training, which is a major limitation of these models. Additionally, these models



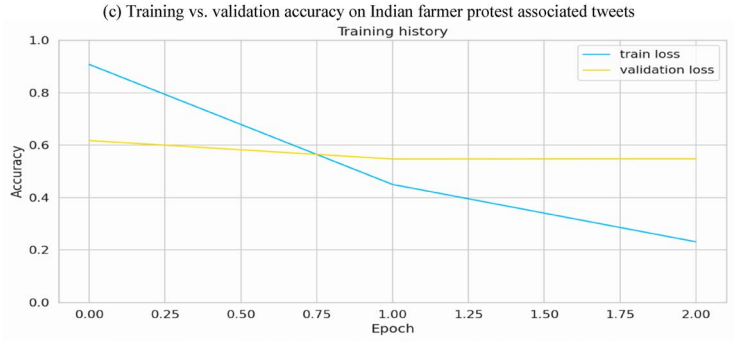
(a) Training vs. validation accuracy on COVID-19 vaccine-associated tweets



(b) Training vs. validation loss on COVID-19 vaccine-associated tweets

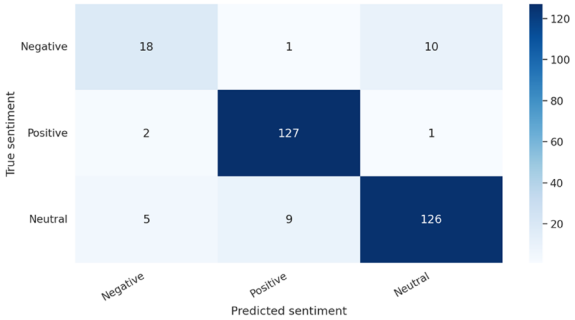


(c) Training vs. validation accuracy on Indian farmer protest associated tweets

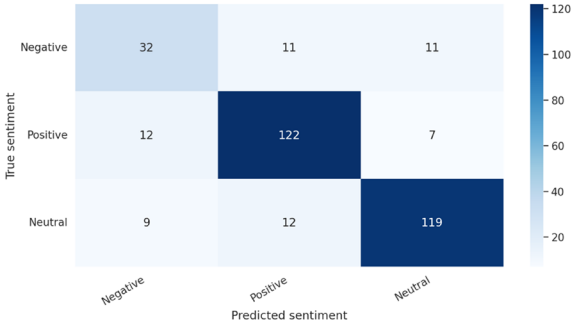


(d) Training vs. validation loss on Indian farmer protest associated tweets

Fig. 12 Training history of the proposed KEAHT model



(a) Confusion matrix calculated by the proposed KEAHT model for COVID-19 vaccine-associated tweets



(b) Confusion matrix calculated by the proposed KEAHT model for Indian farmer protests associated tweets

Fig. 13 Confusion matrix calculated by the proposed KEAHT model

Table 17 The classification report of the proposed KEAHT model

COVID-19 Vaccine-associated tweets				
Sentiment	Precision	Recall	F1-Score	Support
Positive	0.93	0.98	0.95	130
Neutral	0.92	0.90	0.91	140
Negative	0.72	0.62	0.67	29
Micro-average	0.86	0.83	0.84	299
Weighted-average	0.90	0.91	0.90	299
Indian farmer protest-associated tweets				
Positive	0.84	0.87	0.85	141
Neutral	0.87	0.85	0.86	140
Negative	0.60	0.59	0.60	54
Micro-average	0.77	0.77	0.77	335
Weighted-average	0.81	0.81	0.81	335

cannot focus on specific aspects of the complex text input, which misleads the direction of correct sentiment scoring. Consequently, we employ the attention-based mechanism that continuously back-propagates and generates reinforcement learning for effective text learning. Demonstrations reveal that RNN LSTM produces higher results (82% accuracy) than all the experimented machine

Table 18 Comparative results of the existing and proposed KEAHT model

Dataset	Approach	Model	Accuracy	Precision	Recall	F1-Score
COVID-19 vaccine-associated tweets	Machine Learning	Ada Boost	78%	80	78	76
		Extra Tree	80%	80	80	79
		Gradient-Boost	71%	72	71	67
		Random Forest	80%	81	80	78
		Extreme-Gradient Boosting	77%	83	77	75
		Light GBM	69%	67	69	67
	Deep-Learning	Sequential RELU	72%	74	72	73
		CNN	82%	67	82	74
		RNN LSTM	82%	68	81	74
	Transfer Learning	BERT-BASIC	76%	70	77	73
	Hybrid	KEAHT-Proposed	**91%**	**90**	**91**	**90**
Indian farmer protest-associated tweets	Machine Learning	Ada Boost	70%	76	70	67
		Extra Tree	75%	75	75	74
		Gradient-Boost	59%	76	59	51
		Random Forest	76%	77	76	75
		Extreme-Gradient Boosting	71%	76	71	67
		Light GBM	66%	64	66	63
	Deep-Learning	Sequential RELU	68%	65	66	65
		CNN	65%	70	65	67
		RNN LSTM	67%	78	67	54
	Transfer Learning	BERT-BASIC	76%	77	77	77
	Hybrid	KEAHT-Proposed	**81%**	**81**	**81**	**81**

ensembles and deep neural models on the COVID-19 vaccine-associated dataset. Nevertheless, the proposed hybrid model generates approx. 9% more results than superior RNN LSTM and obtains 91.49% accuracy for sentiment classification on the COVID-19 dataset. In the case of Indian farmer's protests associated tweets, Random Forest produces accuracy = 76%, precision = 77, recall = 76, and f1-score = 75, which is the highest of all machine ensembles and deep neural models but lower than the proposed hybrid model. Here, the submitted model scores 5% extra accuracy, 3% extra precision, 4% extra recall, and 6% extra f1-score than the superior Random Forest ensemble. After investigating machine learning and deep learning, the transfer learning-based basic BERT model has also investigated raw tweets about the COVID-19 vaccine and the Indian farmer protest. It has been observed that the proposed KEAHT model acquires 15% and 5% more accuracy than the basic BERT model for COVID-19 vaccine and Indian farmer protest-related raw tweets, respectively. Experiments affirm the superiority of the proposed model that obtains higher results than all the evaluated baseline methods with their inherited potentials.

6.2 Comparison with Existing Hybrid Models

Several researchers have been working in this field and consistently introducing hybrid models with enhanced capabilities. These researches motivated us to develop a novel knowledge-enriched hybrid model to fight against COVID-19 and to support the Indian farmer protests. We compared our hybrid solution with a few existing benchmark models to prove the originality and contribution of our research in the field of NLP. These models were evaluated on multiple datasets; we selected their highest accuracy score and compared them with the higher accuracy score of the proposed model. Table 19 presents the outline of existing state-of-the-art hybrid sentiment analysis models introduced by the researchers in past years.

Figure 14 presents the comparative accuracy score of the existing and proposed hybrid model. The given graph manifests the outperformance of the introduced KEAHT model. Sentiment analysis is the growing field of NLP, and it is utilized to solve various real-life problems. Therefore, it has attracted researchers to launch powerful hybrid models. On the grounds, we selected two critical challenges of real life that not only harmed the few countries, instead of affecting the whole world on a wide scale. Moreover, we developed the hybrid model by adding the knowledge-enriched layer in the attention-based neural model to provide artificial intelligent-based support in these critical situations. Comparative results yielded the achievement of the proposed model over existing baselines and recent state-of-the-art hybrid solutions.

7 Conclusion

The social platforms have allowed us to share our thoughts regarding censorious affairs and help to fight against them by providing the right directions. As everyone knew, the COVID-19 and Indian farmer protests have arisen as two important global concerns that harmed humankind mentally, physically, and economically in the past two years. Meanwhile, social platforms have played a crucial role and joined the various countries' people and governments together for making the right decisions. Social media are meaningless without the support of AI. The intelligent techniques are working behind for mining accurate information from bulky reviews and tweets. The importance of AI for handling the epidemic and democratic dispute motivates us to build a robust hybrid sentiment model. Although several approaches are available for the sentiment classification task, they still have several limitations to correctly identifying the polarity score. Hence, we built a KEAHT hybrid model with the potential of lexicon-based ontology and an attention-based neural model, which can handle the complexity of text inputs. A lexicon-based semantic dictionary has been utilized to generate the tweets' polarity score.

Moreover, the LDA topic modeling has been employed to extract the major topics and their corresponding words, and then these have semantically represented by the NetworkX graphs. This explicit knowledge has fed into the attention network for enriching the process of sentiment classification. Finally, an attention model has

Table 19 Description of existing state-of-the-art hybrid models of sentiment analysis

Author	Model	Method
(Asghar et al. 2017) [59]	T-SAF	The Twitter-Sentiment Analysis Framework (T-SAF) is a hybrid model proposed for classifying the tweets using an emoticon, slang, and domain-specific SentiWordNet classifier
(Zainuddin et al. 2017) [60]	ABSA + SentiWord-Net + PCA + SVM	The complete Aspect-based Sentiment Analysis (ABSA) hybrid model has been proposed with the joint capability of SentiWordNet, Principle Component Analysis (PCA), and SVM
(Ma et al. 2018) [15]	Sentic-LSTM	An explicit knowledge enhances Sentic-LSTM hybrid model was proposed for targeted aspect-based sentiment analysis
(Liu et al. 2019) [61]	AttDR-2DCNN,	An Attention-based Bidirectional and two-Dimensional Convolutional Neural Network (AttDR-2DCNN) was introduced for document-level sentiment classification
(Du et al. 2019) [62]	BGRU + Capsule	A Bi-GRU, a capsule-based hybrid model, has been proposed with the implicit semantic calculation facility
(Meskele[1] and Fransincar, 2019) [46]	ALDONA	A Lexicalized Domain Ontology and Neural Attention Model (ALDONA) has been introduced to handle the multiple polarity aspects in text classification
(Meskele[2] and Fransincar, 2020) [63]	ALDONAr	A Lexicalized Domain Ontology and Regularized Neural Attention Model (ALDONAr), built for aspect-based sentence-level sentiment analysis
(Pathak et al. 2021) [64]	TLSA	A Topic-Level Sentiment Analysis (TLSA) model has been proposed for extracting the opinion of people regarding different cryptocurrencies using the deep learning approach

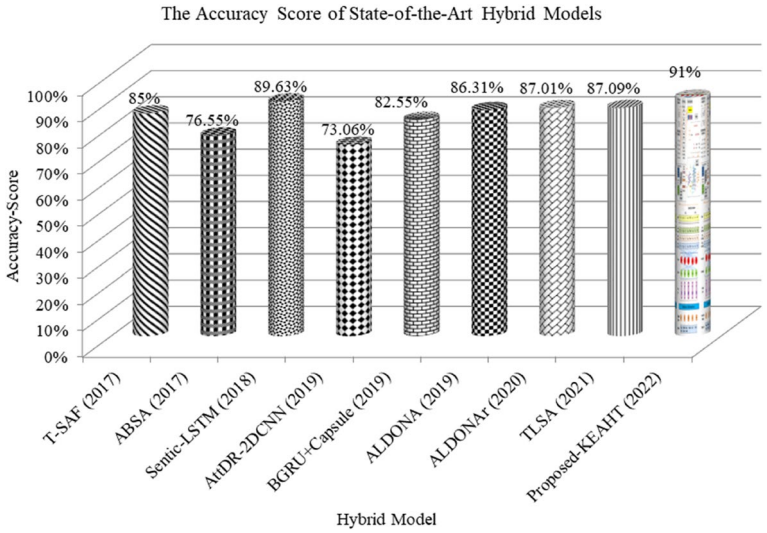


Fig. 14 Comparative performance (Accuracy) of the existing state-of-the-art and proposed KEAHT hybrid model

been trained to mimic a text's cognitive recognition, enhancing the utility part of the input and vanishing the rest for a deeper understanding of the context. This model employs semantic word orientation, unsupervised aspect extraction, extensive BERT word embedding, an attention-based transformer, and an advanced weight decay optimizer to classify the complex tweets. The proposed model has been designed to manipulate the complex formation of the sentence. Unlike DNN, the proposed KEAHT model can train on a small range of datasets with maximum efficiency. An extensive comparison with ensemble baseline methods and seven recently introduced hybrid solutions affirms the credibility of the proposed KEAHT model. Along with this, we developed two benchmark datasets, namely "COVID-19-Vaccine-Labelled-Tweets" and "Indian-Farmer-Protest-Labelled-Tweets," that will be helpful for future researchers to deal with the circumstances mentioned above. Additionally, the various perspectives on the COVID-19 vaccine and Indian farmer protest have been discussed with different handy graphs and images. The proposed model outperformed and showed its efficiency in two different real-life domains. It can also process the long relationships between words in a sentence. This research has few limitations regarding optimal feature selection and statistical analysis, which recommend a promising extension to this research. To contribute to the next-level sentiment model, we will employ advanced feature engineering methods and statistical models like Autoregressive Integrated Moving Average (ARIMA) with deep learning techniques that will enhance the capability of sentiment classification applications. Furthermore, the extension of lexicalized domain ontology would better capture the relationships of new concepts that can lead to more reliability. Thus, we will investigate the semi-automatic approach to enhance the domain ontology for the sentiment analysis context.

Data availability The data that support the findings of this study are openly available at: COVID-Vaccine Tweets: https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets Indian-Farmer Protest Tweets: https://github.com/AlbusDracoSam/Farmers-protest-tweet-analysis/blob/main/tweetsfarmbills.csv

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Phan, H.T., Tran, V.C., Nguyen, N.T., Hwang, D.: Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. IEEE Access **8**, 14630–14641 (2020)
2. Sailunaz, K., Alhajj, R.: Emotion and sentiment analysis from Twitter text. J. Comput. Sci. **36**, 101003 (2019)
3. Alamoudi, A.H., Zaidan, B.B., Zaidan, A.A., Albahri, O.S., Mohammed, K.I., Malik, R.Q., Alaa, M.: Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review. Expert Syst. Appl. **167**, 114155 (2021)
4. Neogi, A.S., Garg, K.A., Mishra, R.K., Dwivedi, Y.K.: Sentiment analysis and classification of Indian farmers' protest using twitter data. Int. J Inf. Manag. Data Insights **1**(2), 100019 (2021)
5. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
6. Balahur, A., Turchi, M.: Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput. Speech Lang. **28**(1), 56–75 (2014)
7. da Silva, N.F.F., Coletta, L.F., Hruschka, E.R., Hruschka, E.R., Jr.: Using unsupervised information to improve semi-supervised tweet sentiment classification. Inf. Sci. **355**, 348–365 (2016)
8. Souma, W., Vodenska, I., Aoyama, H.: Enhanced news sentiment analysis using deep learning methods. J. Comput. Soc. Sci. **2**(1), 33–46 (2019)
9. Chaturvedi, I., Ragusa, E., Gastaldo, P., Zunino, R., Cambria, E.: Bayesian network-based extreme learning machine for subjectivity detection. J. Franklin Inst. **355**(4), 1780–1797 (2018)
10. Basiri, M.E., Nemati, S., Abdar, M., Cambria, E., Acharya, U.R.: ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis. Futur. Gener. Comput. Syst. **115**, 279–294 (2021)
11. Lauren, P., Qu, G., Yang, J., Watta, P., Huang, G.B., Lendasse, A.: Generating word embeddings from an extreme learning machine for sentiment analysis and sequence labeling tasks. Cogn. Comput. **10**(4), 625–638 (2018)
12. Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., Tian, G.: Incorporating word embeddings into topic modeling of short text. Knowl. Inf. Syst. **61**(2), 1123–1145 (2019)
13. Ishaq, A., Asghar, S., Gillani, S.A.: Aspect-based sentiment analysis using a hybridized approach based on CNN and GA. IEEE Access **8**, 135499–135512 (2020)
14. Usama, M., Ahmad, B., Song, E., Hossain, M.S., Alrashoud, M., Muhammad, G.: Attention-based sentiment analysis using convolutional and recurrent neural network. Futur. Gener. Comput. Syst. **113**, 571–578 (2020)
15. Ma, Y., Peng, H., Khan, T., Cambria, E., Hussain, A.: Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cogn. Comput. **10**(4), 639–650 (2018)
16. Minh, D.L., Sadeghi-Niaraki, A., Huy, H.D., Min, K., Moon, H.: Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. Ieee Access **6**, 55392–55404 (2018)
17. Setiawan, E.I., Ferry, F., Santoso, J., Sumpeno, S., Fujisawa, K., Purnomo, M.H.: Bidirectional GRU for targeted aspect-based sentiment analysis based on character-enhanced token-embedding and multi-level attention. Computing **1**, 2 (2020)
18. Zhou, J., Lu, Y., Dai, H.N., Wang, H., Xiao, H.: Sentiment analysis of Chinese microblog based on stacked bidirectional LSTM. IEEE Access **7**, 38856–38866 (2019)

19. Gupta, N., Agrawal, R.: Application and techniques of opinion mining. In: Hybrid computational intelligence, pp. 1–23. Academic Press (2020)
20. Darwich, M., Mohd, S.A., Omar, N., Osman, N.A.: Corpus-based techniques for sentiment lexicon generation: a review. J. Digit. Inf. Manag. **17**(5), 296 (2019)
21. Abdulla, N.A., Ahmed, N.A., Shehab, M.A., Al-Ayyoub, M., Al-Kabi, M.N., Al-rifai, S.: Towards improving the lexicon-based approach for Arabic sentiment analysis. Int J Inf Technol Web Eng (IJITWE) **9**(3), 55–71 (2014)
22. Saif, H., He, Y., Fernandez, M., Alani, H.: Contextual semantics for sentiment analysis of Twitter. Inf. Process. Manage. **52**(1), 5–19 (2016)
23. Khoo, C.S., Johnkhan, S.B.: Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons. J. Inf. Sci. **44**(4), 491–511 (2018)
24. Kundi, F.M., Khan, A., Ahmad, S., Asghar, M.Z.: Lexicon-based sentiment analysis in the social web. J. Basic Appl Sci Res **4**(6), 238–248 (2014)
25. Keshavarz, H., Abadeh, M.S.: ALGA: adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. Knowl.-Based Syst. **122**, 1–16 (2017)
26. Kumar, A., Jain, R.: Uniform textual feedback analysis for effective sentiment analysis. In: Iberoamerican knowledge graphs and semantic web conference, pp. 273–289. Springer, Cham (2021)
27. Le, L., Patterson, A., White, M.: Supervised autoencoders: improving generalization performance with unsupervised regularizers. Adv. Neural. Inf. Process. Syst. **31**, 107–117 (2018)
28. Park, S., Lee, J., Kim, K.: Semi-supervised distributed representations of documents for sentiment analysis. Neural Netw. **119**, 139–150 (2019)
29. Rintyarna, B.S., Sarno, R., Fatichah, C.: Evaluating the performance of sentence-level features and domain sensitive features of product reviews on supervised sentiment analysis tasks. J Big Data **6**(1), 1–19 (2019)
30. Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., González-Castaño, F.J.: Unsupervised method for sentiment analysis in online texts. Expert Syst. Appl. **58**, 57–75 (2016)
31. Unnisa, M., Ameen, A., Raziuddin, S.: Opinion mining on Twitter data using unsupervised learning technique. Int J Comput Appl **148**(12), 975–8887 (2016)
32. Kumar, N., Venugopal, D., Qiu, L., Kumar, S.: Detecting anomalous online reviewers: an unsupervised approach using mixture models. J. Manag. Inf. Syst. **36**(4), 1313–1346 (2019)
33. Khan, F.H., Qamar, U., Bashir, S.: A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. Knowl. Inf. Syst. **51**(3), 851–872 (2017)
34. Khan, F.H., Qamar, U., Bashir, S.: SWIMS: semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis. Knowledge-Based Syst. **100**, 97–111 (2016)
35. Rani, S., Kumar, P.: Deep learning-based sentiment analysis using convolution neural network. Arab. J. Sci. Eng. **44**(4), 3305–3314 (2019)
36. Jain, P.K., Saravanan, V., Pamula, R.: A hybrid CNN-LSTM: a deep learning approach for consumer sentiment analysis using qualitative user-generated contents. Trans. Asian Low-Resour Lang Inf Process **20**(5), 1–15 (2021)
37. Rehman, A.U., Malik, A.K., Raza, B., Ali, W.: A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimedia Tools and Applications **78**(18), 26597–26613 (2019)
38. Daval-Frerot, G., Bouchekif, A., & Moreau, A.: Epita at SemEval-2018 task 1: sentiment analysis using transfer learning approach. In: Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 151–155. (2018)
39. Singh, M., Jakhar, A.K., Pandey, S.: Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc. Netw. Anal. Min. **11**(1), 1–11 (2021)
40. Gujjar, J.P., Kumar, H.P.: Sentiment analysis: Textblob for decision making. Int. J. Sci. Res. Eng. Trends **2021**, 1097–1099 (2021)
41. Zhang, H., Gan, W., Jiang, B.: Machine learning and lexicon-based methods for sentiment classification: a survey. In: 2014 11th web information system and application conference, pp. 262–265 (2014)
42. Jurek, A., Mulvenna, M.D., Bi, Y.: Improved lexicon-based sentiment analysis for social media analytics. Secur. Inf. **4**(1), 1–13 (2015)
43. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools Appl. **78**(11), 15169–15211 (2019)

44. Song, Y., Pan, S., Liu, S., Zhou, M. X., Qian, W.: Topic and keyword re-ranking for LDA-based topic modeling. In: Proceedings of the 18th ACM conference on information and knowledge management, pp. 1757–1760 (2009)
45. Tiwari, D., Nagpal, B.: Ensemble sentiment model: bagging with linear discriminant analysis (BLDA). In: 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 474–480, IEEE (2021)
46. Meškelė, D., Frasincar, F.: ALDONA: a hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalised domain ontology and a neural attention model. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 2489–2496 (2019)
47. Sharma, M., Ilanthenral, K., Vasantha, W.B.: Comparison of neutrosophic approach to various deep learning models for sentiment analysis. Knowledge-Based Syst. **223**, 107058 (2021)
48. Bhati, B.S., Chugh, G., Al-Turjman, F., Bhati, N.S.: An improved ensemble-based intrusion detection technique using XGBoost. Transac. Emerg. Telecommun. Technol. **32**(6), e4076 (2021)
49. Vizcarra, J., Kozaki, K., Torres Ruiz, M., Quintero, R.: Knowledge-based sentiment analysis and visualization on social networks. New Gener. Comput. **39**(1), 199–229 (2021)
50. Tiwari, D., Singh, N.: Ensemble approach for twitter sentiment analysis. IJ Inform. Technol. Comput. Sci. **11**, 20–26 (2019)
51. Adhikari, N. C. D., Kurva, V. K., Suhas, S., Kushwaha, J. K., Nayak, A. K., Nayak, S. K., Shaj, V.: Sentiment classifier and analysis for epidemic prediction. SAI, ICAITA, CSITA, ISPR, Signal, 31–48 (2018)
52. Bansal, A., Kumar, N.: Aspect-based sentiment analysis using attribute extraction of hospital reviews. New Gener. Comput. **27**, 1–20 (2021)
53. Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kwak, K.S.: Transportation sentiment analysis using word embedding and ontology-based topic modeling. Knowledge-Based Syst. **174**, 27–42 (2019)
54. Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Adam, S.: Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun. Methods Meas. **12**(2–3), 93–118 (2018)
55. Roussille, P., Megdiche, I., Teste, O., Trojahn, C.: Boosting holistic ontology matching: generating graph clique-based relaxed reference alignments for holistic evaluation. In: European Knowledge Acquisition Workshop (pp. 355–369). Springer, Cham (2018)
56. Wang, M., Lu, S., Zhu, D., Lin, J., Wang, Z.: A high-speed and low-complexity architecture for softmax function in deep learning. In: 2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp. 223–226 (2018)
57. Adam Optimizer: https://ruder.io/optimizing-gradient-descent/index.html#adam, Accessed 23 Dec 2021
58. Tiwari, D., Bhati, B.S., Nagpal, B., Sankhwar, S., Al-Turjman, F.: An enhanced intelligent model: to protect marine IoT sensor environment using ensemble machine learning approach. Ocean Eng. **242**, 110180 (2021)
59. Asghar, M.Z., Kundi, F.M., Ahmad, S., Khan, A., Khan, F.: T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. Expert. Syst. **35**(1), e12233 (2018)
60. Zainuddin, N., Selamat, A., Ibrahim, R.: Hybrid sentiment classification on twitter aspect-based sentiment analysis. Appl. Intell. **48**(5), 1218–1232 (2018)
61. Liu, F., Zheng, J., Zheng, L., Chen, C.: Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. Neurocomputing **371**, 39–50 (2020)
62. Du, Y., Zhao, X., He, M., Guo, W.: A novel capsule-based hybrid neural network for sentiment classification. IEEE Access **7**, 39321–39328 (2019)
63. Meškelė, D., Frasincar, F.: ALDONAr: a hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. Inf. Process. Manag. **57**(3), 102211 (2020)
64. Pathak, A.R., Pandey, M., Rautaray, S.: Topic-level sentiment analysis of social media data using deep learning. Appl. Soft Comput. **1**(108), 107440 (2021)

Authors and Affiliations

Dimple Tiwari[1] · **Bharti Nagpal**[2]

[1] Research Scholar, Ambedkar Institute of Advanced Communication Technologies and Research (GGSIPU), New Delhi, India

[2] NSUT East Campus (Formerly Ambedkar Institute of Advanced Communication Technologies and Research), New Delhi, India