Social Network Analysis and Mining (2023) 13:57
https://doi.org/10.1007/s13278-023-01066-z

ORIGINAL ARTICLE

Deep learning-based credibility conversation detection approaches from social network

Imen Fadhli[1] · Lobna Hlaoua[1] · Mohamed Nazih Omri[1]

Received: 15 February 2023 / Revised: 10 March 2023 / Accepted: 11 March 2023 / Published online: 29 March 2023

Abstract
In recent years, the social networks that have become most exploited sources of information, such as Facebook, Instagram, LinkedIn, and Twitter, have been considered the main sources of non-credible information. False information on these social networks has a negative impact on the credibility of conversations. In this article, we propose a new deep learning-based credibility conversation detection approach in social network environments, called CreCDA. CreCDA is based on: (i) the combination of post and user features in order to detect credible and non-credible conversations; (ii) the integration of multi-dense layers to represent features more deeply and to improve the results; (iii) sentiment calculation based on the aggregation of tweets. In order to study the performance of our approach, we have used the standard PHEME dataset. We compared our approach with the main approaches we have studied in the literature. The results of this evaluation show the effectiveness of sentiment analysis and the combination of text and user levels to analyze conversation credibility. We recorded the mean precision of credible and non-credible conversations at 79%, the mean recall at 79%, the mean F1-score at 79%, the mean accuracy at 81%, and the mean G-Mean at 79%.

Keywords Credibility detection · Twitter conversation · Post features · User features · Deep learning · Sentiment analysis

1 Introduction

1.1 Context and issue

The evolution of social media has changed people's attention from newspapers toward popular social media platforms Corradini et al. (2021). Social media, such as Facebook and Twitter, have now created a new field of information. Twitter is one of the most popular social media, with 500 million tweets sent per day and 199 million users in 2021[1]. This massive collection of tweets contains users' opinions and information, such as events and interests, feelings, personal ideas, etc., Alrubaian et al. (2018). Furthermore, with this new technology, a large number of information sources have emerged and become more accessible. As a result, the control and filter mechanisms become more difficult. The authors in Gammoudi et al. (2022) presented a literature review of influencer identification on social media. Fake news and rumors are one of the principal influences on user credibility Gammoudi et al. (2022). The evaluation of the credibility of sources, messages, and media represents an important topic to study Metzger et al. (2003). Spammers use the anonymity features of microblogs to distribute fake news and spam messages via scam URLs Qureshi et al. (2021). The COVID-19 pandemic is an important example that demonstrates the impact of misinformation on social media. The introduction of COVID-19 vaccines in 2020 has created various claims on social media, particularly on Twitter Goodman and Carmichael (2020). The famous claim was: "Bill Gates, the co-founder of Microsoft, plans to implant trackable microchips into vaccine doses in order to control and geo-locate people." Goodman and Carmichael (2020). Following the BBC's announcement of Bill

Lobna Hlaoua and Mohamed Nazih Omri have contributed equally to this work.

✉ Imen Fadhli
imen.fdh@gmail.com

Lobna Hlaoua
lobna1511@yahoo.fr

Mohamed Nazih Omri
Mohamednazih.omri@eniso.u-sousse.tn

[1] MARS Research Laboratory LR17ES05, University of Sousse, Sousse, Tunisia

[1] https://www.omnicoreagency.com/twitter-statistics/.

and Melinda Gates, this claim was identified as fake news Goodman and Carmichael (2020). Furthermore, COVID-19 of vaccines has created new discussions on Twitter in which users were divided into two groups: those who believe in vaccines, called pro-vaxxers and those who doubt their efficacy, called anti-vaxxers. Bonifazi et al. (2022). The propagation of incorrect and/or false news influences the public's perception of a critical topic. In addition, they presented a significant risk to global health and indirectly influenced the worsening of conditions. So, it is critical to assess how the exchange of content on social networks influences the perception of such situations, which is one of the major purposes of the social network analysis area of research.

Misinformation on social media is a difficult task, particularly when it is related to sensitive topics not only healthcare but also politics. According to the study of Allcott and Gentzkow (2017), the disseminated fake news on Twitter during the 2016 American presidential elections had a significant impact on voters. In El Ballouli et al. (2017), the authors discovered that roughly 40% of daily tweets are untrustworthy. Moreover, the authors of Gupta et al. (2013) presented a study of the propagation of false information during Hurricane Sandy. They discovered that 86% of misinformation were retweeted. They demonstrated that during a crisis, people share news from untrustworthy source. So, in an emergency situation, analyzing the credibility of Twitter is a crucial issue. Information credibility has recently become a major concern due to the increased volume of news and information on Twitter. Furthermore, the short length of tweets (280 characters) has contributed to the large and uncontrolled number of tweets. Unfortunately, the presence of fake sources, noise, and misinformation has an influence on the quality of tweet content. We concentrate on Twitter conversations in order to improve the quality of tweets Ahmad (2022). Several authors have proposed various methods for calculating the credibility of Twitter content. Some researchers employ machine learning techniques, while others employ graph-based techniques and human perception judgments Ahmad (2022) Al-Khalifa and Al-Eidan (2011).

1.2 Contributions

A conversation is a collection of tweets and replies exchanged between users. The most significant difference between conversation credibility and tweet credibility is the existence of interactions between users. The communication produced by a user's comments is crucial to classifying a conversation as credible or not. Our main contributions to the analysis of conversation credibility are as follows. (i) We calculated conversation credibility by combining "textual features" and "user features." (ii) After using the CNN-LSTM model, we integrate different dense layers to represent features more deeply. (iii) Using the CSAM model Fadhli et al. (2022), we calculated sentiment conversation by aggregating tweets comments. We also conducted tests to confirm their effectiveness.

1.3 Paper organization

The rest of our article is organized as follows: in Sect. 2, we provide an overview of the credibility conversation detection approach from social network. Section 3 presents the Problem formulation. Section 4 presents our proposed CreCDA Approach. Section 5 carries out an experimental study of our proposed approach in order to verify its effectiveness and compare it with the main approaches that we have studied in the context of this article. Section 6 mentions the limits of our proposed model. In Sect. 7, we conclude our work and highlight some ideas for future work.

2 Related work

The authors of Qureshi et al. (2021) defined credibility as "believability, trust, reliability, accuracy, fairness, and objectivity." According to the Oxford dictionary, credibility is "the quality that someone/something possesses that causes people to believe or trust them." Several research studies have identified two aspects of credibility: expertise and trustworthiness. The literature presents various levels of credibility:

- Post level: Analyze tweets' content to calculate a credibility score based on diverse features such as tweets containing hashtags (#), @ mentions, or links; verbs, and nouns; multimedia content (video, image, audio) features; sentiment features (positive, negative words) Castillo et al. (2011). It is calculated through text analysis techniques, such as Natural Language Processing (NLP) Zubiaga et al. (2017), Hassan et al. (2020), Azer et al. (2021), Yamaguchi et al. (2010).
- User level: measures credibility from features of the user's account, for example, the user's age, number of followers, friends, tweets, and retweets Azer et al. (2021), Abbasi and Liu (2013), Al-Sharawnh et al. (2013), Yamaguchi et al. (2010).
- Topic level: determines the topic or event's credibility. It identifies topics by combining URLs, hashtags, the number of verbs or nouns that describe the event, etc., generally using NLP and sentiment analysis techniques Abu-Salih et al. (2019).
- Hybrid level: This level combines the post, topic, and user levels to calculate credibility Castillo et al. (2011); Abu-Salih et al. (2019).

There are several methods for classifying credible and untrustworthy tweets. Recent research divides credibility methods into three categories: Automatics-based approaches, human-based approaches, and hybrid approach Alrubaian et al. (2018) Qureshi et al. (2021).

2.1 Credibility evaluation

2.1.1 Automatics-based approaches

Many studies use automated methods to assess the credibility of tweets, such as machine learning, graph-based methods, and weighted algorithms.

Machine learning: machine learning algorithms were developed to analyze information that includes large volumes of data with various parameters. There are two types of techniques that fit into this category:

- Supervised techniques: such as support vector machine (SVM), logistic/linear regression models, Bayesian theory, and decision tree methods Castillo et al. (2011); Zubiaga et al. (2017); Hassan et al. (2020); Abu-Salih et al. (2019); Azer et al. (2021); Qureshi and Sabih (2021); Giachanou et al. (2021); Zhang et al. (2021); Ouni et al. (2022); Omri and Omri (2022).
- Unsupervised techniques: including clustering models (e.g., k-means, fuzzy c-means, or hidden Markov models) Abbasi and Liu (2013); Al-Sharawnh et al. (2013); Gangireddy et al. (2020).

Many supervised techniques are used in the literature to assess trustworthiness. The first to investigate Twitter's credibility were Castillo et al. (2011). They demonstrate that the combination of J48 decision tree, SVM, and Bayesian approaches with different features can be highly indicative of credibility content. Their methods, Castillo et al. (2011), have a credibility accuracy of 86%. In Zubiaga et al. (2017), the authors developed a new method for detecting rumors on social media. The authors of Zubiaga et al. (2017) compared various machine learning techniques for rumor analysis. Similar work has also been pursued by others Hassan et al. (2020) in which they proposed a classification model based on supervised machine methods with word-N-gram features to classify tweets as credible or not credible. In Abu-Salih et al. (2019), the authors created a new framework, "CredSat," to assess user credibility. They assessed credibility by combining post, topic, and user features with domain-based credibility. In more recent work Azer et al. (2021), the authors used machine learning algorithms to discover rumors from tweets. They achieved an accuracy of 83.44% Azer et al. (2021), by combining two types of features (post and user level) and a random forest classifier. Furthermore, the work Giachanou et al. (2021) presented emoCred model that is based on a long-short term memory (LSTM) model that incorporates emotional signals extracted from the text of the claims to distinguish between credible and non-credible text in social networks. The authors of Zhang et al. (2021) proposed a model based on the interactions between speakers in a conversation: (i) in adding a confidence gate before each LSTM hidden unit to estimate the credibility of the previous speakers, and (ii) in combining the output gate with the learned influence scores to incorporate the influences of the previous speakers. Similarly to supervised methods, there exists a very extensive literature on this topic. This issue was explored by the works in Abbasi and Liu (2013) and Al-Sharawnh et al. (2013). For example, in the first works, the authors developed a novel method called CredRank to analyze user behavior in social networks and rank their trustworthiness. This model was created to group related users and count the number of people in each group. Clustering compares how users behave in various social networks. Tweet, for example, is clustered based on the similarity of their users' tweets. According to the work proposed in Al-Sharawnh et al. (2013), knowledge is disseminated on Twitter by powerful people known as "leaders" or "pioneers." They proposed a method for assessing leaders' trustworthiness by analyzing their social network impressions and participation in events such as crises. In Ouni et al. (2022), the authors combined BERT (bidirectional encoder representations of transformers) and CNN (convolutional neural network) to detect spammers on social networks.

Graph-based methods: In the paper, Hamdi et al. (2020), the authors proposed a method to determining the credibility of information sources on Twitter. They used the node2vec algorithm to extract features from the Twitter follower graph. The authors of Yamaguchi et al. (2010) proposed TURank (Twitter User Rank), an algorithm based on link analysis for determining users' authority scores on Twitter. Another work is presented in Gupta et al. (2012) where the authors presented a credibility analysis approach that includes event graph-based optimization to assign similar scores to similar events in order to overcome the problem of rumors on Twitter. In Gangireddy et al. (2020), the authors created GTUT to demonstrate the usage of graph mining methods over the textual, user, and temporal data from social networks to distinguish between fake and real news on Twitter.

Weighted Algorithms: Researchers used this set of approaches to determine the trustworthiness of the source and user information. In Al-Khalifa and Al-Eidan (2011), the authors proposed a model of similarity between tweets, authentic, and URLs. Another work presented in Widyantoro and Wibisono (2014) shows that the tweets were represented using TF-IDF as well as the cosine function, which is the most commonly used method for calculating topic similarity and weighting terms. The proposed approach in Middleton (2015) developed a project called REVEAL to detect false and credible messages using a set of regex patterns matching both terms and POS tags to determine media credibility. This project analyzes tweets at the post level, which is insufficient for determining the credibility of a tweet. It is also based on well-known journalistic verification principles, which limits its adaptability.

2.1.2 Human-based approaches

Human-based approaches employ statistical analysis tools such as questionnaires, interviews, etc. The work presented in Kawabe et al. (2015) realized that several factors, including the user's influence and topical expertise, improve credibility. They also increase the importance of the message topic, which influences people's perceptions of tweet credibility, with science-related topics receiving the highest rating, followed by politics and entertainment. Another approach presented in Qiu et al. (2014) discovered that tweets that began with "@"(indicating direct responses to other users) were less common during such events. It has also been established that during times of crisis, the number of tweets containing URLs decreases.

2.1.3 Hybrid approaches

Hybrid approaches combine both automatic and human-based approaches to analyze the credibility of social media content Kim and Hastak (2018); Jaho (2014). For example, the authors in Jaho (2014) used a clustering approach with weighted and IR algorithms. The model, presented in Ito et al. (2015), combined topic and user features with latent Dirichlet allocation (LDA), clustering, and a random forest classifier to assess information credibility.

The majority of existing works on social media credibility identification leaned on manually extracting features and addresses a gap in the sentiment analysis of conversations. Our approach is similar to the research presented in Ahmad (2022). It combines the post and user levels to analyze the credibility of Twitter conversations while ignoring the sentiment of each conversation. Using a deep learning framework, we present a novel set of features for determining credibility on Twitter. In addition, to analyze sentiment in conversation, we use a novel sentiment analysis model.

2.2 Tweet credibility and sentiment analysis

Sentiment analysis has been used in various social network research projects, particularly on Twitter. Many frameworks have been developed to evaluate the trustworthiness of user content in the context of social trust, taking into account users' sentiments about the content of their tweets. The model proposed in Abu-Salih et al. (2019) considers not only the sentiment of tweets but also the sentiment of replies. Similar works detailed in Zubiaga et al. (2017) and Hassan et al. (2020) have also been carried out by the work presented in Azer et al. (2021) in which they integrated sentiment features in tweet content to analyze fake news. The authors of the approach presented in Azer et al. (2021) obtained the best results by combining sentiment analysis features with user and post features using random forest (RF). The authors of Castillo et al. (2011) proposed another model that examined the most important aspects of the credibility task. They show that the presence of positive and negative terms in tweets increases the content's credibility.

As part of this work, in the rest of this paper, we will examine credibility conversations at the post and user levels. However, measuring the credibility of tweets at this level is difficult, not only because of the character limit (280 characters) but also because single tweets do not express and provide information to determine the event or theme. Furthermore, the presence of a positive tweet in the conversation does not guarantee that the conversation is positive. Thus, the following section describes the problem formulation of the conversation credibility detection problem from the social network.

3 Problem formulation

This problem can be described as a binary classification, which looks as follows. Formally, let C^T be the set of tweets T_i and R_i the set of corresponding replies, with $1 \leq i \leq N$. We model the set of conversations C^T by a combination of tweets, denoted $C^T=\{T_i, R_{i+1}, R_{i+2},..., R_N\}$. T_i represents the initial tweet in C^T. R_i is the i-th tweet reply, in the conversation that is a set of tweets exchanged between users about a specific topic. We consider that a conversation contains not only tweets and responses but also users. So, we represent users' conversation by : $C^U=\{U_i, U_{i+1}, U_{i+2},..., U_N\}$. The conversation content is insufficient to represent the conversation's credibility. Therefore, we choose to combine content with user interest. User social reputation is one of the most important features. We present our approach credibility conversation detection approach (CreCDA) based on a CNN-LSTM neural network that combines post and user levels to differentiate between credible and non-credible conversations. To improve the performance of our model and to learn the features more deeply, we chose to add some dense layers after applying CNN-LSTM. Our goal is to classify conversations as credible or not in terms of tweet content, user comments, and user interactions.

4 Proposed CreCDA approach

In this section, we present our proposed CreCDA approach from social network. We propose its general architecture, and its different preprocessing steps. Then, we present the

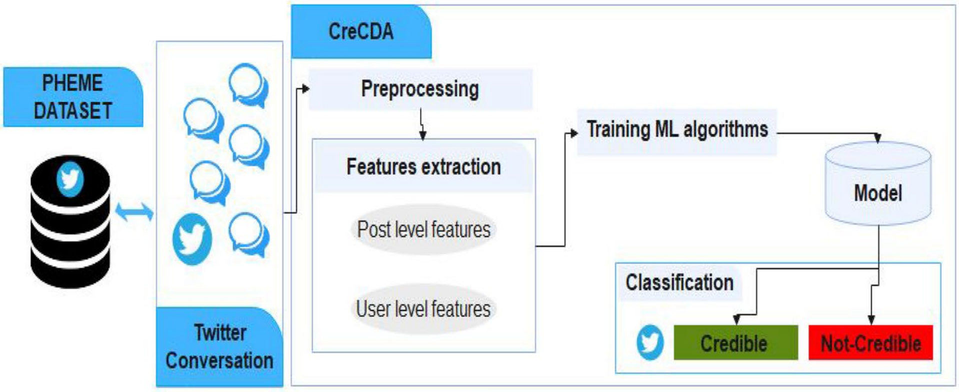


Fig. 1 General architecture of the proposed model

process of features extraction we consider, we detail the training model, and we finish by presenting the CreCDA algorithm we developed.

4.1 Architecture of the proposed model

The flowchart in Figure 1 illustrates the different stages of the model detection process that we have proposed. The dataset extraction phase is the first step in the process. The second step is preprocessing, used to remove noise from the dataset. This step generates spelling and punctuation corrections. The third step is Word2Vec, which will be detailed in the following section. Then, the features extraction stage, in which we select the relevant features for credibility tasks. Finally, transfer those features to a credibility prediction machine learning model.

4.2 Preprocessing steps of CreCDA model

Preprocessing data is critical, especially for social media content. Because of their short length (280 characters), incomplete abbreviations, and expressions, tweets are well known as unstructured and noisy data. These issues have a negative impact on the detection of credibility. To reduce noise in tweets, the following preprocessing steps are used:

- Remove punctuation: the punctuation marks such as commas, apostrophes, and question marks, which do not have much importance to the credibility detection process, are removed.
- Numbers removal step: when measuring credibility, the numbers are useless and in order to improve the content of the tweets, the numbers are removed.
- Remove Stopwords: In this step, stopwords such as “and, that, and a” are removed.
- Lowercasing: one of the fundamental cleaning operations that convert a word to lowercase.
- Tokenization: is the process of working with substance data to confine a bit of substance into small units called tokens. The tokens are tallying segments and sentences, which can be broken into words with assistance.

4.3 Features extraction

The process of selecting and extracting relevant information from given data and forming feature vectors is called feature extraction. There are many methods for converting text to vector; in this work, we used word embedding which is a popular method of representing the vocabulary of a document in which words with similar meanings are represented similarly. Word2vec is a well-known neural network model for learning word embeddings. It takes a text as input and generates a low-dimensional vector of the words in the text corpus. Several features to measure the reliability of news events on Twitter have been proposed in the literature. For the purposes of this article, we consider one of the most useful and popular feature sets available today. The work illustrated in Sharma (2019) suggests three factors to consider when determining credible information: (i) source of information, (ii) content of information, and (iii) comments of the user. In order to analyze credibility conversations, we focus on publication levels. The following sections describe all the features offered.

4.3.1 Post-level features

Post-level or content-based features are the features that focus on the substance of the conversation that may influence the credibility of the conversation.

Conversation-content: In this research, we measure the conversation content by cosine similarity. Cosine similarity is a metric used to determine the similarity of documents, irrespective of their size. We measure the cosine similarity between the source tweet and the reaction tweet.

Sentiment conversation: Several studies have attempted to improve the significance of emojis and emoticons in detecting sentiment, including Park et al. (2018); Azer et al. (2021). Before calculating the polarity of tweets, we convert emojis and emoticons ":) :(:|" to textual data, so ":)" will be converted to “happy face smiley.” To convert these emojis and emoticons, we use the Python library. The polarity of each tweet in the conversation is then calculated. In this step, we propose to use two models to calculate sentiment conversation. The following section details the proposed models to calculate sentiment conversation.

CreDeep sentiment conversation: In the step of calculating the polarity score, we use the TextBlob sentiment analysis model with Twitter conversations Silva et al. (2020). TextBlob Silva et al. (2020): TextBlob is a rule-based sentiment analysis library that focuses on lexical content and integrates sentiment analysis with the WordNet corpus. The calculated polarity score ranges [−1...1]. (−1 negative sentiment, 1 positive sentiment, 0 for neutral). We calculate the sentiment score of a conversation with the model given by equation 1 according to Abu-Salih et al. (2019).

$$\text{Senti}(C^T) = \frac{\text{Pol}(T_0) - \min(\text{Pol}(R^*))}{\max(\text{Pol}(R^*)) - \min(\text{Pol}(R^*))} \tag{1}$$

With: $\text{Pol}(T_0)$: polarity of the initial tweet using TextBlob. $\min(\text{Pol}(R^*))$: represents the minimum polarity of replies in the conversation. $\max(\text{Pol}(R^*))$: represents the maximum polarity of replies in the conversation. The main issue with this model is that it ignores the dependencies between tweets and comments, it also ignores the variation in polarity in the conversation. So, in order to solve this problem, we propose to use CSAM model Fadhli et al. (2022).

CSAM sentiment conversation: We assume that if we have a positive response in the conversation, it does not confirm that the conversation is positive, and the same is true for negative tweets. Indeed, CSAM consists of calculating a degree of belief using the theory of evidence to identify the meaning of the sentimental conversation. This degree of belief, denoted by *bel*, measures the total belief attributed to a conversation.

$$\text{Bel}(C^T) = \sum_{i=1}^{N} m(c_i) = \sum_{i=1}^{N} \text{Prob}_i(R_i|Ti) \tag{2}$$

$m(c_i)$ is a conditional probability computed by combining polarity with similarity.

$$\text{Prob}(T|R) = \frac{\text{Pol}(T) * \text{sim}_{\cos}(T,R)}{\text{Pol}(R)} \tag{3}$$

T represents a tweet and R is a reply.

Pol: score polarity calculated by *Textblob* sentiment analysis model.

Cosine similarity: Represents the most commonly used measure of similarity between two vectors, which is based on the co-occurrence of terms by fixing the dimension on the different concepts. Similarity is an important factor in detecting credible conversations Choudhary et al. (2018). The similarities between the initial tweet and replies are more critical in influencing trust and understanding than those that are perceived to be dissimilar Schouten et al. (2020). The similarity is defined as follows:

$$\text{Sim}_{\cos}(T,R) = \frac{w2vec(T) * w2vec(R)}{||w2vec(T)|| * ||w2vec(R)||} \tag{4}$$

Where $w2vec(T)$, $w2vec(R)$ are two vectors calculated by *word2vec* techniques. $Sim_{cos(T,R)}$ is within the range [−1, 1], the value −1 indicates that the vectors are opposite, 0 for independent vectors and 1 for similar.

Table 1 Proposed features

Levels	
Post level	User level
Conversation-content	User-verified
Sentiment conversation	Ratio-followers
	Ratio-friends
	Ratio-status
	Rank-user

4.3.2 User-level features

It is well known that the interaction of users is a significant factor when discovering socially reliable content. This involves studying followers' and friends' interests and the users' content, and so on. At this level, we focus on the characteristics of the author (the source) of the tweet as shown in Table 1.

User-verified: As to whether the current user account is verified or non-verified, 1 represents a verified account and 0 represents a non-verified account.

Ratio-followers: According to Azer et al. (2021) calculated as the number of followers divided by the account age. First, we calculate the ratio of followers per user to check if it was a fake account or not. Then, we determine the ratio of followers by conversation, in which we divide the obtained ratio by the number of tweets in the conversation.

$$\text{Ratio}_{\text{Flw}}(U_i) = \frac{\text{Nb}_{\text{Flw}}(U_i)}{\text{account}_{\text{age}}(U_i)} \tag{5}$$

$$\text{Ratio}_{\text{Flw-conv}}(C^U) = \frac{\sum \text{Ratio}_{\text{Flw}}(U_i)}{\text{Nb}_{\text{user}}(C^U)} \tag{6}$$

Ratio-friends: To start, we calculate the number of friends divided by the account age Azer et al. (2021). Then, we calculate the ratio of friends in conversation with equation 5.

$$\mathrm{Ratio}_{\mathrm{Frd}}(U_i) = \frac{\mathrm{Nb}_{\mathrm{Frd}}(U_i)}{\mathrm{account}_{\mathrm{age}}(U_i)} \tag{7}$$

$$\mathrm{Ratio}_{\mathrm{Frd-conv}}(C^U) = \frac{\sum \mathrm{Ratio}_{\mathrm{Frd}}(U_i)}{\mathrm{Nb}_{\mathrm{user}}(C^U)} \tag{8}$$

Ratio-status: We begin by calculating the number of statuses per user divided by the account age Azer et al. (2021). Then, we calculate the ratio of status in conversation with equation 7.

$$\mathrm{Ratio}_{\mathrm{Sts}}(U_i) = \frac{\mathrm{Nb}_{\mathrm{Sts}}(U_i)}{\mathrm{account}_{\mathrm{age}}(U_i)} \tag{9}$$

$$\mathrm{Ratio}_{\mathrm{Sts-conv}}(C^U) = \frac{\sum \mathrm{Ratio}_{\mathrm{Sts}}(U_i)}{\mathrm{Nb}_{\mathrm{user}}(C^U)} \tag{10}$$

Rank-user: The gratitude of users' conversation increases when the user receives a lot of consideration. According to Ahmad (2022) rank of a user can be expressed as follows:

$$\mathrm{Rank}_{\mathrm{user}}(U_i) = -\log \frac{\sum_{i=1}^{N} \mathrm{favorite}_i + \mathrm{comment}_i + \mathrm{retweet}_i}{\mathrm{follower}_i} \tag{11}$$

We consider a conversation is a set of users. We define the rank of a user per conversation as given by equation 9.

$$\mathrm{Rank}_{\mathrm{user-conv}}(C^U) = \frac{\sum \mathrm{Rank}_{\mathrm{user}}(U_i)}{\mathrm{Nb}_{\mathrm{user}}(C^U)} \tag{12}$$

4.4 Training model

In our research on credible conversation detection, we used the CNN-LSTM model. To make our detection model more efficient and deep optimized, we used a multi-dense layer. The number of dense layers, activation functions, and loss functions are as follows: The first step after feature extraction is to represent conversation content and user features in a separate embedding layer. The first layer in the model is called the input layer, which contains the input features. We have two input layers, one for post features and another for user features. It has a dimension of (32, 300), where 32 represents the batch size and 300 is the input vector length. The second layer is the embedding layer. Embeddings are an excellent method for NLP issues for two reasons. First, it reduces dimensionality over one-hot encoding because we can control the number of features. Second, because similar words have similar embedding, the embedding layer can understand a word's context. Next, we apply the CNN-LSTM model, in which each feature will be assigned an activation function and passed on to the next layer in the network. After that, we combine the output layer obtained after the application of the CNN-LSTM model. To improve the performance of our model and to learn features more deeply, we propose to add four dense layers: dense(10), dense(8), dense(6), and dense(4), respectively.

Dense layer A dense layer's functionality can be defined as a linear operation in which every input is connected to every output by some weight. The addition of hidden layers in the neural network improves the model to a point, but adding more layers can harm the model's performance (it depends upon the complexity of the problem). Our proposed model has four dense layers to achieve the best results. Researchers in Alrubaian et al. (2021) have mostly used one or two dense layers.

Activation function As an activation method, we used ReLU (Rectified Linear Unit). The main reason for using ReLU is that it successfully removes negative values from an activation map in a network by setting them to zero. It is the most commonly used activation function in deep learning. It solves the vanishing gradient problem more efficiently than sigmoid or Tanh. It is 0 for all negative values of input z and equal to z for all positive values of input z. The ReLU equation is as follows:

$$\sigma = \max(0, z) \tag{13}$$

Loss function The final step is the output layer, in which we use a cross-entropy function. The performance of a classification model whose output is a probability value between 0 and 1 is measured by cross-entropy loss, also known as log loss. As the predicted probability diverges from the actual label, cross-entropy loss increases. Cross-entropy can be calculated in binary classification when the number of classes equals 2.

$$\mathrm{Loss} = -(y\log(p) + (1-y)\log(1-p)) \tag{14}$$

The layered architecture of our multi-dense credibility based model (CreCDA) is shown Fig. 2 and Table 2. This neural network-based model has 300 input nodes. It has four hidden layers. The first dense layer contains 10 nodes. There are 8 hidden nodes in the second dense layer. The third layer contains 6 hidden nodes. Fourth dense layer has 4 hidden nodes. The final output layer has 2 nodes and the sigmoid activation function.

The CNN-LSTM architecture used in our experiments is given by the following equations:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{15}$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{16}$$

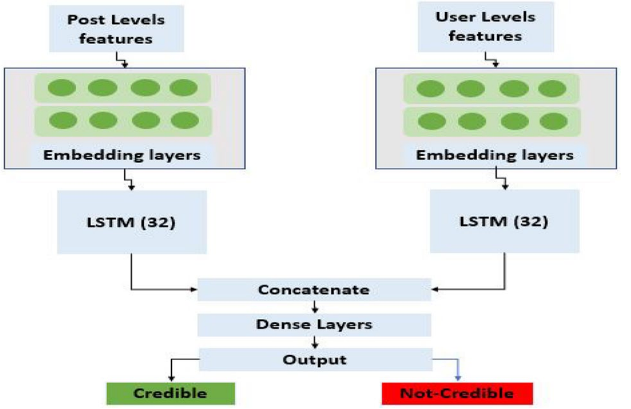


Fig. 2 Preprocessing steps of CreCDA

Table 2 Layered architecture of our proposed CreCDA model

Layer	Input	Output
Dense	64	10
Dense_1	10	8
Dense_2	8	6
Dense_3	6	4
Dense_4	4	2

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{17}$$

i_t: represents the input gates. f_t: represents the forget gates. o_t: represents the output gates. σ : sigmoid function.

We compared the performance of six different classifiers, namely: logistic regression (LR), random forests (RF), convolution neural network (CNN), convolution neural network–long short-term memory (CNN-LSTM), and long short–term memory (LSTM) in order to choose the best classifier.

4.5 Proposed algorithm

Algorithm 1 Proposed CreCDA algorithm

```
Input: Conv = {$C_0^T$, ..., $C_N^T$}
$C^T$ = {$T_0$, ..., $R_N$}
Output: Credible Conversation
begin
    for i in $C^T$ do
        for j in range (i+1,$C^T$) do
            /* Calculation similarity
            $Sim_{cos}(T_i, R_j)$ [13]
        end
    end
    Calculate sentiment analysis: CSAM [13]
    foreach $T_i$ in $C^T$ do
        Extract users features (eq3),(eq5),(eq7),(eq9)
    end
    /* Initialize LSTM parameters
    Training iteration: Itr=164
    Training batch size: B=32
    Training epochs: E=10
    Randomly initialize weight: w
    /* Training
    for [1...Itr] do
        Calculate the output: y=LSTM (eq12)
        Calculate the loss function : (eq13)
        Calculate weights
        Backpropagation and update w value

    end
    /* Testing
    calculate predicted value: y with testing set
end
End
```

5 Experimental study and results analysis

In this section, we present the setup used to implement our approach (Sect. 5.1), the datasets used to approach the task of credible conversation (Sect. 5.2), and the different metrics used to evaluate the performance of our model (Sect. 5.3). Section 5.4 outlines the results of the study. Finally, the findings are discussed (Sect. 5.5).

5.1 Simulation setup

The proposed solution was implemented in Python 3.10 language using Jupyter notebook 6.4.11 on the platform anaconda 3. The simulations are carried out on a PC Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz 2.40 GHz, 6 GB RAM. Table 3 represent hyperparameters used for our proposed deep neural network-based model "CreCDA."

Deep neural network-based model CreCDA: CreCDA with four layers 10,8,6,4 is designed. ReLU for hidden layers and sigmoid for the output layer is used as activation function. Adam optimizer is used for optimizing the designed CreCDA. CreCDA is trained for 10 epochs.

Table 3 Hyperparameters for our proposed CreCDA model

Hyperparameter	Description or value
No. of dense layers	5
No. of hidden nodes	32,32,64,10,8,6,4,2
Activation function	ReLU
Loss function	Cross-entropy
Output layer	Sigmoid
Number of epochs	10
Batch-size	32
Optimizer	Adam

5.2 Data collection

We used the PHEME dataset, which is the most commonly used dataset in the credibility detection task Zubiaga et al. (2017). The dataset was collected by Zubiaga et al. (2017) using the Twitter streaming API during five breaking news events. Table 4 presents all events connected to the PHEME dataset as well as the total number of credible and non-credible. The total number of tweets is 5802, which were manually annotated as 3830 (66%) credible and 1972 (34%) non-credible tweets. This dataset is labeled dataset. One represents a rumor, and zero represents a not-rumor. In our case, we classified rumors as non-credible and not-rumors as credible (Table 4).

5.3 Evaluation metrics

To evaluate our approach, we used four common measures, frequently used for assessing the performance of the binary classifiers. The confusion matrix in Table 5 presents these measures, which are defined as follows:

Table 4 PHEME Dataset base

Event dataset	Credible	Not-credible	Total
Charlie Hebdo	458	1621	2079
Ferguson	284	859	1143
German wings crash	238	231	469
Ottawa shooting	470	420	890
Sydney siege	522	699	1221
Total	1972	3830	5802

Table 5 Confusion matrix

	Predictive result	
Actual result	TP	FN
	FP	TN

- TP (True Positives): is the number of positive conversations detected correctly classified as "positive,"
- TN (True negatives): is the number of negative conversations detected correctly classified as "negative,"
- FP (False positives): is the number of negative conversations detected incorrectly classified as "positive," and
- FN (False negatives): is the number of positive conversations detected incorrectly classified as "negative."

To calculate the performance of our model, we used the following metrics:

- Precision: represents the percentage of true positive instances is predicted to be positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{18}$$

- Recall: the recall is the proportion of true positive instances from cases that are actually positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

- F1-score: this measure represents the average of the precision and recall.

$$\text{F1} - \text{score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{20}$$

- Accuracy: the subset of the predicted real instances compared to the set of predicted instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{21}$$

- G-mean: indicates the geometric mean of the recall.

$$\text{G} - \text{mean} = \sqrt{\frac{\text{TN}}{\text{TN} + \text{FP}} * \frac{\text{TP}}{\text{TP} + \text{FN}}} \tag{22}$$

5.4 Experimental results and analysis

We have tested our model with five ML algorithms: LR, RF, CNN, CNN-LSTM, and LSTM. The results of this analysis are summarized in Tables 6, 7, 8, 9, 10, 11, 12 and 13. We have applied our model to the PHEME dataset Zubiaga et al. (2015) using precision, recall, F1-score, accuracy, and G-mean as the performance metrics. Firstly, we tested our model with post features as shown in Tables 6, 7, and Fig. 3. We obtained the best results with credible conversation in terms of precision with LSTM (0.66), recall with CNN model (0.75), F1 with CNN-LSTM model (0.66). We add two metrics to confirm our result's accuracy and G-mean and we obtained the best score with CNN-LSTM (0.76) accuracy

Table 6 The score results of credible conversation from five trained models of the proposed approach "CreDeep" using post features

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CreDeep_{LR}$	0.11	0.01	0.01	0.64	0.07
$CreDeep_{RF}$	0.52	0.06	0.1	0.66	0.23
$CreDeep_{CNN}$	0.57	0.75	0.65	0.72	0.69
$CreDeep_{LSTM}$	0.66	0.57	0.61	0.75	0.71
$CreDeep_{CNN-LSTM}$	**0.65**	**0.68**	**0.66**	**0.76**	**0.71**

Bold values indicate the obtained results with CNN-LSTM model

Table 7 The score results of non-credible conversation from five trained models of the proposed approach "CreDeep" using post features

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CreDeep_{LR}$	0.65	0.98	0.78	0.64	0.07
$CreDeep_{RF}$	0.66	0.97	0.79	0.66	0.23
$CreDeep_{CNN}$	0.84	0.71	0.77	0.72	0.69
$CreDeep_{LSTM}$	0.79	0.85	0.82	0.75	0.71
$CreDeep_{CNN-LSTM}$	**0.83**	**0.81**	**0.82**	**0.76**	**0.71**

Table 8 The score results of credible conversation from five trained models of the proposed approach "CSAM" using post features

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CSAM_{LR}$	0.4	0.05	0.09	0.65	0.21
$CSAM_{RF}$	0.55	0.03	0.06	0.66	0.17
$CSAM_{CNN}$	0.74	0.76	0.75	0.82	0.8
$CSAM_{LSTM}$	0.63	0.72	0.67	0.76	0.74
$CSAM_{CNN-LSTM}$	**0.77**	**0.61**	**0.69**	**0.81**	**0.74**

Table 9 The score results of non-credible conversation from five trained models of the proposed approach "CSAM" using post features

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CSAM_{LR}$	0.66	0.96	0.78	0.65	0.21
$CSAM_{RF}$	0.66	0.99	0.79	0.66	0.17
$CSAM_{CNN}$	0.87	0.86	0.87	0.82	0.8
$CSAM_{LSTM}$	0.84	0.78	0.81	0.76	0.74
$CSAM_{CNN-LSTM}$	**0.82**	**0.91**	**0.86**	**0.81**	**0.74**

and (0.71) with G-mean. In the scenario of a non-credible conversation, we note that the best recall rate (0.98) but with a precision rate of (0.65) was obtained by applying the LR learning model. The CNN learning model resulted in the highest precision rate (0.84) but a recall rate of (0.71). We noticed that applying the learning models LSTM and CNN-LSTM yielded the best values of F1-score (0.82). We calculate accuracy and the G-mean score we obtained (0.76) with CNN-LSTM and (0.71) with CNN-LSTM and LSTM models.

Secondly, we calculate sentiment analysis with CSAM model Fadhli et al. (2022) and we obtained the results in Tables 8, 9, and Fig. 4. The best results of credible conversation are generated with the CNN model in terms of recall, F1-score, accuracy, and G-mean (0.76), (0.75), (0.82), and (0.8), respectively, while the best precision with the CNN-LSTM learning model was (0.77). In the case of non-credible conversation, the highest recall with the RF model was (0.99), whereas the best precision, F1-score, accuracy, and g-mean with the CNN model. The results show that there is only a minor difference in score between CNN and CNN-LSTM (Fig. 7).

Third, we combine conversation content with user features as in Tables 10,11, and Fig. 5. We achieved the best recall, F1 measure, and G-mean in credible conversation with the CNN-LSTM model at (0.69), (0.65), and (0.74), respectively, while the highest precision and accuracy with the LSTM model (0.7) and (0.77). On the other hand, with non-credible conversation, we obtained a higher recall with LR and RF (0.97). However, with the LSTM model we achieved (0.87) F1-score and (0.83) of accuracy, whereas the best precision and G-mean with the CNN-LSTM model were (0.83) and (0.74), respectively. Values in bold indicate the best results.

Finally, to improve the results, we combine the "CreDeep" model with "CSAM." Tables 12, 13, and Fig. 6 summarize the findings of this analysis. The table shows that the best results are achieved with both CNN and CNN-LSTM for credible and non-credible.

As shown in Tables 10,11,12, and 13, we observe a remarkable increase with the combination of the CSAM and CreDeep models. When there is a credible conversation, the increase in scores ranges between [0.05] and [0.18]. The obtained results increased from [0.02] to [0.09] for a non-credible conversation with both the CNN

Table 10 The score results of credible conversation from five trained models of the proposed approach "CreDeep" using post+user features

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CreDeep_{LR}$	0.5	0.06	0.1	0.66	0.23
$CreDeep_{RF}$	0.5	0.06	0.1	0.66	0.23
$CreDeep_{CNN}$	0.64	0.59	0.61	0.74	0.71
$CreDeep_{LSTM}$	0.7	0.57	0.63	0.77	0.7
$CreDeep_{CNN-LSTM}$	**0.61**	**0.69**	**0.65**	**0.74**	**0.74**

Table 11 The score results of non-credible conversation from five trained models of the proposed approach "CreDeep" using post+user features

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CreDeep_{LR}$	0.66	0.97	0.79	0.66	0.23
$CreDeep_{RF}$	0.66	0.97	0.79	0.66	0.23
$CreDeep_{CNN}$	0.79	0.82	0.81	0.74	0.71
$CreDeep_{LSTM}$	0.79	0.87	0.83	0.77	0.7
$CreDeep_{CNN-LSTM}$	**0.83**	**0.77**	**0.8**	**0.74**	**0.74**

Table 12 The score results of credible conversation from five trained models of the proposed approach "CreCDA" using CreDeep post+user features and CSAM model

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CreCDA_{LR}$	0.5	0.04	0.07	0.66	0.18
$CreCDA_{RF}$	0.5	0.04	0.07	0.66	0.18
$CreCDA_{CNN}$	0.71	0.77	0.74	0.81	0.8
$CreCDA_{LSTM}$	0.65	0.77	0.71	0.78	0.77
$CreCDA_{CNN-LSTM}$	**0.71**	**0.75**	**0.73**	**0.81**	**0.79**

Table 13 The score results of non-credible conversation from five trained models of the proposed approach "CreCDA" using CreDeep post+user features and CSAM model

Approaches	Metrics				
	Precision	Recall	F1-score	Accuracy	G-mean
$CreCDA_{LR}$	0.66	0.98	0.79	0.66	0.18
$CreCDA_{RF}$	0.66	0.98	0.79	0.66	0.18
$CreCDA_{CNN}$	0.87	0.84	0.85	0.81	0.8
$CreCDA_{LSTM}$	0.87	0.78	0.82	0.78	0.77
$CreCDA_{CNN-LSTM}$	**0.87**	**0.84**	**0.85**	**0.81**	**0.79**

and CNN-LSTM models, while the LSTM model shows a slight decrease in precision and accuracy for credible conversations, as well as in recall, F1-score accuracy, and G-mean for non-credible conversations. Tables 12 and 13 summarize the obtained results, we achieve the best results with CNN and CNN-LSTM models for credible conversation seems as non-credible. The highest precision score (0.71) for credible and (0.87) for non-credible. The best recall, F1-score, and accuracy were achieved with CNN for credible conversation. We obtained the highest recall of not credible conversations with LR and RF (0.98). However, our model achieved the best recall, F1-score, and accuracy with CNN and CNN-LSTM models (Table 13).

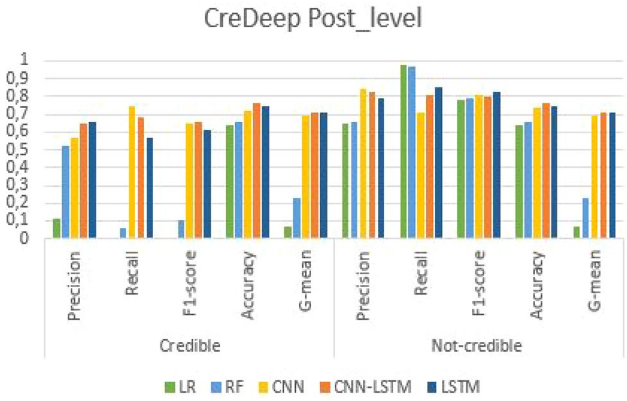


Fig. 3 The score results of "CreDeep" using post+user features

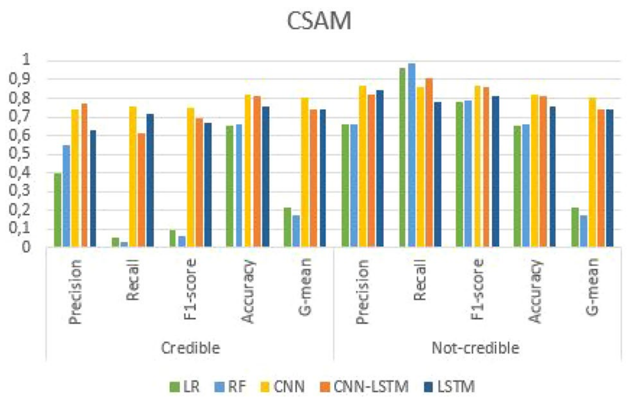


Fig. 4 The score results of m"CSAM" using post features

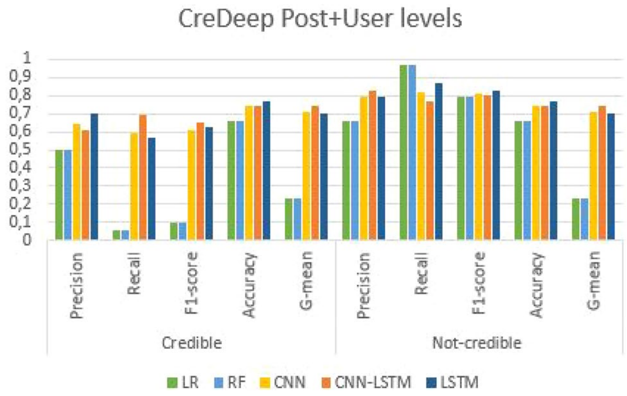


Fig. 5 The score results of "CreDeep"using post+user features and CNN-LSTM model

We observe that we obtained close results between CNN and CNN-LSTM. Figure 7 shows the CNN/CNN-LSTM accuracy and loss curves after training versus the number of epochs for the classification task based on combined content and social context. The curves show that the model has

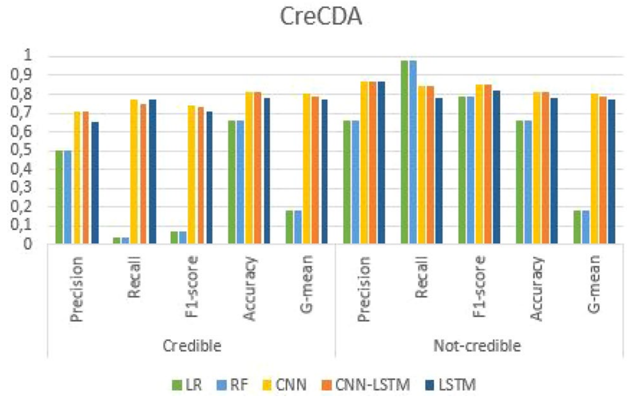


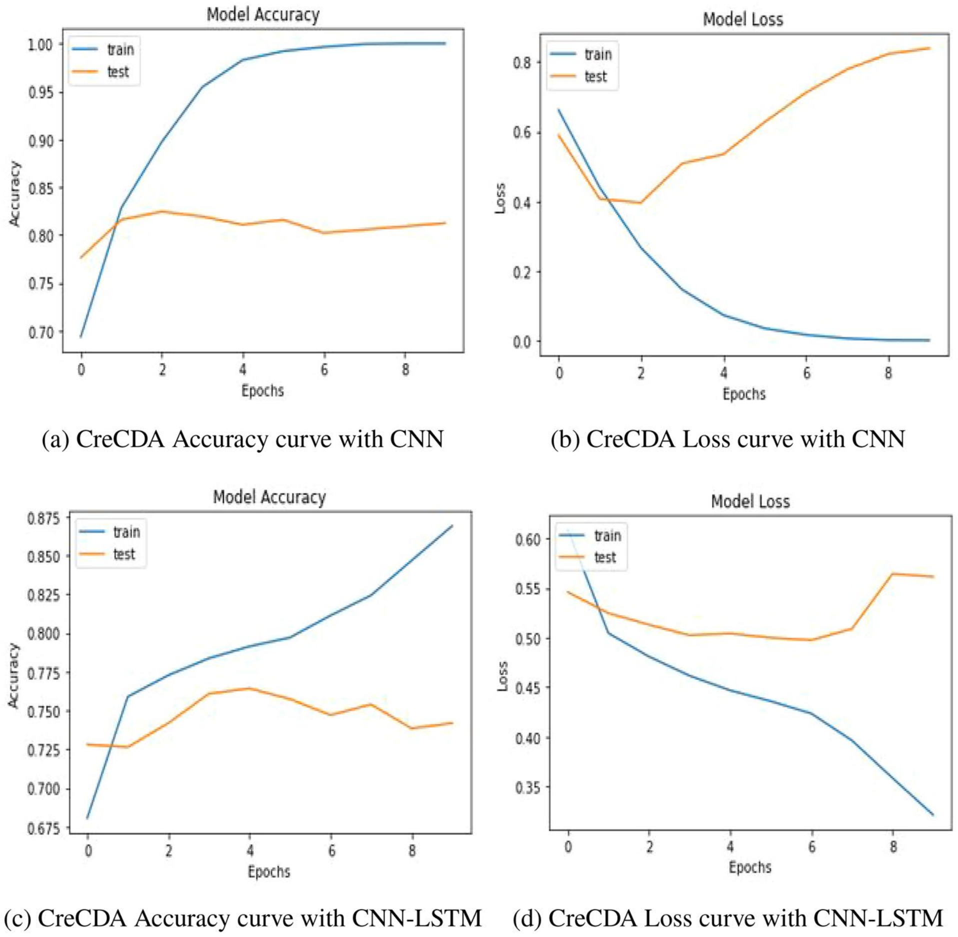
Fig. 6 The score results of the proposed approach "CreCDA" post+user features and CNN-LSTM model

learned with CNN-LSTM better than with CNN. Furthermore, Fig. 7b and d from 7 show that CreCDA overfits the CNN model, with a loss value of (0.83) versus (0.6) for the CNN-LSTM. We conclude that CreCDA with CNN-LSTM improves the results compared to the CNN model. We recorded that our model increases recall. This means that our model is able to detect non-credible conversations in the corpus. Table 14 and Fig. 8 summarizes the results of our model when compared to the CSAM and CreDeep models.

5.5 Discussion

PHEME corpus has been used in several studies, including Zubiaga et al. (2015, 2017); Kotteti et al. (2018). They use deep learning to analyze credible conversations. On the PHEME dataset, our model is compared to the existing state-of-the-art models. To begin, as shown in Table 15 we compare our model with Zubiaga et al. (2017); Kotteti et al. (2018); Bharti (2020) to determine whether tweets were rumors or not. We discovered that our model performed best with non-credible conversations. This demonstrates that our model detects untrustworthy conversations more effectively than trustworthy ones. We calculated the mean precision, recall, and F1 between credible and non-credible conversations, as shown in Table 15. Second, we compare our model to Kotteti et al. (2018); Zubiaga et al. (2017) which classified Twitter conversations as rumors and

Fig. 7 CreCDA accuracy and loss curve with combined Post+user features and CNN-LSTM/CNN model



(a) CreCDA Accuracy curve with CNN

(b) CreCDA Loss curve with CNN

(c) CreCDA Accuracy curve with CNN-LSTM

(d) CreCDA Loss curve with CNN-LSTM

Table 14 Scores of credible and non-credible conversations for various models

Credibility	Approach	Metrics				
		Precision	Recall	F1-score	Accuracy	G-mean
Credible	CSAM	**0.77**	0.61	0.69	**0.81**	0.74
	CreDeep	0.61	0.69	0.65	0.74	0.74
	Our approach CreCDA	0.71	**0.75**	**0.73**	**0.81**	**0.79**
Non-credible	CSAM	0.82	**0.91**	**0.86**	**0.81**	0.74
	CreDeep	0.83	0.77	0.8	0.74	0.74
	Our approach CreCDA	**0.87**	0.84	0.85	**0.81**	**0.79**

not-rumors. In Kotteti et al. (2018), the authors used time series data to reduce time and supervised learning algorithms. On the other hand, Zubiaga et al. (2017) used a conditional random field (CRF) based on content and social features. They Kotteti et al. (2018) had the highest precision of (0.94), while our model had the second-highest precision of (0.79). When compared to Kotteti et al. (2018); Zubiaga et al. (2017), our model achieved the best recall and F1 measures among all classes. The results show that these methods can produce very good classification results. We conclude that the use of the "CSAM" sentiment analysis aggregation model and the combination of textual and user levels improve the results. Furthermore, this approach can be helpful to improve the performance of credible conversation detection. The use of our proposed CreCDA further improves the performance as compared to both CreDeep and CSAM.

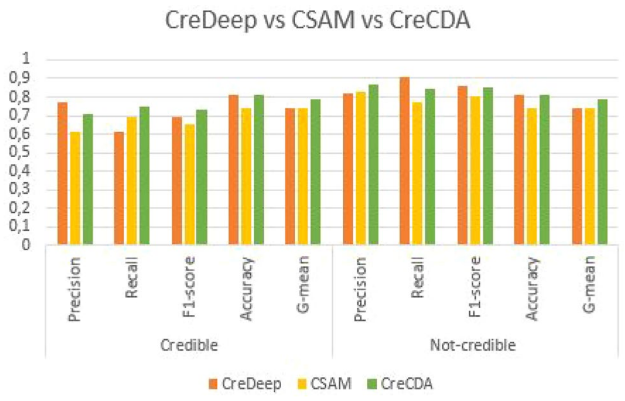


Fig. 8 The score results of the proposed approach using post+user features and LR model

6 Limits of the proposed approach

The main goal of the proposed approach is to propose an efficient solution for a credible Twitter conversation. Despite its high precision, recall, f-measure, G-mean, and accuracy, our model ignores the semantic relationships between terms used to represent information and the relationships between these concepts. Furthermore, it does not integrate hashtags as microblog features. This limitation penalizes our solution's excellent performance and prevents us from considering it a perfect solution. This strategy could be used in the future.

7 Conclusion and future works

7.1 Summary

In this paper, we have presented a new model to detect credible conversations. First, we have used a new model of calculating sentiment conversation based on a belief function and conditional probability. Second, calculate conversation credibility by combining post and user features based on the CNN-LSTM model. To evaluate our method, we used the PHEME dataset. First, we test our model with post features. We use five ML models. The results show that the use of CNN-LSTM with our model outperforms CNN, LSTM, LR, and RF. Second, we evaluated our model with post-user features, in which we achieved high scores with the CNN-LSTM model. We have conducted different standard measures such as precision, recall, F1-score, accuracy, and G-mean. The obtained results confirm the feasibility of our model and its performance. We also compared our proposed approach with other models in the literature like CSAM and CreDeep.

7.2 Prospects

Our future work can be articulated in three directions. As a first direction, we plan to evaluate our model with recent conversation datasets in order to compare it to state-of-the-art approaches. The second direction is to use different semantic features with the contextual representation of words and thematic features. The integration of these new characteristics and their consideration will probably improve the performance of our model on credible searches on

Table 15 Comparison of the proposed approach "CreCDA" with Bharti (2020), Zubiaga et al. (2017) and Kotteti et al. (2018) of credible/non-credible conversation using PHEME dataset

Credibility	Approaches	Metrics				
		Precision	Recall	F1-score	Accuracy	G-mean
Credible	Bharti et al Bharti (2020)	**0.87**	**0.89**	**0.88**	–	–
	Our approach CreCDA	0.71	0.75	0.73	**0.81**	**0.79**
Non-credible	Bharti et al Bharti (2020)	0.79	0.76	0.77	–	–
	Our approach CreCDA	**0.87**	**0.84**	**0.85**	**0.81**	**0.79**
All	Zubiaga et al Zubiaga et al. (2017)	0.66	0.55	0.6	–	–
	Kotteti et al Kotteti et al. (2018)	**0.94**	0.35	0.51	–	–
	Our approach CreCDA	0.79	**0.79**	**0.79**	**0.81**	**0.79**

Bold values indicate the best results

Twitter. The third direction consists in using an ontological structure at the post level allowing the representation of the semantic links between the terms and the relations between the concepts for better detection of credible conversations on Twitter.

Author Contributions The authors confirm contribution to the paper as follows: F: study conception and design, data collection, analysis and interpretation of results, draft manuscript preparation: F, NO, and H. All authors reviewed the results and approved the final version of the manuscript.

Funding The authors received no financial support for the research and/or authorship of this article.

Declarations

Conflict of interest The authors declare no competing interests.

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

References

Abbasi M-A, Liu H (2013) Measuring user credibility in social media. In: International conference on social computing, behavioral-cultural modeling, and prediction, Springer. pp 441–448

Abu-Salih B, Wongthongtham P, Chan KY, Zhu D (2019) Credsat: credibility ranking of users in big social data incorporating semantic analysis and temporal factor. J Inf Sci 45(2):259–280

Ahmad T (2022) Efficient fake news detection mechanism using enhanced deep learning model. Appl Sci 12(3):1743

Al-Khalifa HS, Al-Eidan RM (2011) An experimental system for measuring the credibility of news content in twitter. Int J Web Inf Syst

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236

Alrubaian M, Al-Qurishi M, Alamri A, Al-Rakhami M, Hassan MM, Fortino G (2018) Credibility in online social networks: a survey. IEEE Access 7:2828–2855

Alrubaian M, Al-Qurishi M, Omar S, Mostafa, MA (2021) Deeptrust: a deep learning approach for measuring social media users trustworthiness. arXiv preprint arXiv:2101.07725

Al-Sharawnh J, Sinnappan S, Williams M-A (2013) Credibility-based twitter social network analysis. In: Asia-Pacific web conference, Springer. pp 323–331

Azer M, Taha M, Zayed HH, Gadallah M (2021) Credibility detection on twitter news using machine learning approach. Int J Intell Syst Appl 13(3):1–10

Bharti (2020) Automatic rumour detection model on social media. In: 2020 sixth international conference on parallel, distributed and grid computing (PDGC), pp 367–371

Bonifazi G, Breve B, Cirillo S, Corradini E, Virgili L (2022) Investigating the covid-19 vaccine discussions on twitter through a multilayer network-based approach. Inf Process Manag 59(6):103095

Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the 20th international conference on world wide web, pp 675–684

Choudhary N, Singh R, Bindlish I, Shrivastava M (2018) Neural network architecture for credibility assessment of textual claims. arXiv preprint arXiv:1803.10547

Corradini E, Nocera A, Ursino D, Virgili L (2021) Investigating the phenomenon of nsfw posts in reddit. Inf Sci 566:140–164

El Ballouli R, El-Hajj W, Ghandour A, Elbassuoni S, Hajj H, Shaban K (2017) Cat: credibility analysis of arabic content on twitter. In: Proceedings of the third Arabic natural language processing workshop, pp 62–71

Fadhli I, Hlaoua L, Omri MN (2022) Sentiment analysis csam model to discover pertinent conversations in twitter microblogs

Gammoudi F, Sendi M, Omri MN (2022) A survey on social media influence environment and influencers identification. Soc Netw Anal Min 12(1):1–19

Gangireddy SCR, Long C, Chakraborty T (2020) Unsupervised fake news detection: A graph-based approach. In: Proceedings of the 31st ACM conference on hypertext and social media, pp 75–83

Giachanou A, Rosso P, Crestani F (2021) The impact of emotional signals on credibility assessment. J Assoc Inf Sci Technol 72(9):1117–1132

Goodman J, Carmichael F (2020) Coronavirus: Bill gates 'microchip' conspiracy theory and other vaccine claims fact-checked. BBC News **30**

Gupta A, Lamba H, Kumaraguru P, Joshi A (2013) Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on world wide web, pp 729–736

Gupta M, Zhao P, Han J (2012) Evaluating event credibility on twitter. In: Proceedings of the 2012 SIAM international conference on data mining, SIAM. pp 153–164

Hamdi T, Slimi H, Bounhas I, Slimani Y (2020) A hybrid approach for fake news detection in twitter based on user features and graph embedding. In: International conference on distributed computing and internet technology, Springer. pp 266–280

Hassan N, Gomaa W, Khoriba G, Haggag M (2020) Credibility detection in twitter using word n-gram analysis and supervised machine learning techniques. Int J Intell Eng Syst 13(1):291–300

Ito J, Song J, Toda H, Koike Y, Oyama S (2015) Assessment of tweet credibility with lda features. In: Proceedings of the 24th international conference on world wide web, pp 953–958

Jaho (2014) Alethiometer: a framework for assessing trustworthiness and content validity in social media. In: Proceedings of the 23rd international conference on world wide web, pp 749–752

Kawabe T, Namihira Y, Suzuki K, Nara M, Sakurai Y, Tsuruta S, Knauf R (2015) Tweet credibility analysis evaluation by improving sentiment dictionary. In: 2015 IEEE congress on evolutionary computation (CEC), pp 2354–2361

Kim J, Hastak M (2018) Social network analysis: characteristics of online social networks after a disaster. Int J Inf Manag 38(1):86–96

Kotteti CMM, Dong, X, Qian L (2018) Multiple time-series data analysis for rumor detection on social media. In: 2018 IEEE international conference on big data (Big Data), pp 4413–4419

Metzger MJ, Flanagin AJ, Eyal K, Lemus DR, McCann RM (2003) Credibility for the 21st century: integrating perspectives on source, message, and media credibility in the contemporary media environment. Ann Int Commun Assoc 27(1):293–335

Middleton S (2015) Extracting attributed verification and debunking reports from social media: mediaeval-2015 trust and credibility analysis of image and video

Omri MN, Omri F (2022) Dynamic editing distance-based extracting relevant information approach from social networks. Int J Comput Netw Inf Secur

Ouni S, Fkih F, Omri MN (2022) Bert-and cnn-based tobeat approach for unwelcome tweets detection. Soc Netw Anal Min 12(1):1–19

Park, et al (2018) Plusemo2vec at semeval-2018 task 1: exploiting emotion knowledge from emoji and# hashtags. arXiv preprint arXiv:1804.08280

Qiu Q, Xu R, Liu B, Gui L, Zhou Y (2014) Credibility estimation of stock comments based on publisher and information uncertainty evaluation. In: International conference on machine learning and cybernetics, Springer. pp 400–408

Qureshi KA, Sabih M (2021) Un-compromised credibility: Social media based multi-class hate speech classification for text. IEEE Access 9:109465–109477

Qureshi KA, Malick RAS, Sabih M (2021) Social media and microblogs credibility: identification, theory driven framework, and recommendation. IEEE Access 9:137744–137781

Schouten AP, Janssen L, Verspaget M (2020) Celebrity vs. influencer endorsements in advertising: the role of identification, credibility, and product-endorser fit. Int J Advert 39(2):258–281

Sharma K (2019) Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol (TIST) 10(3):1–42

Silva, et al (2020) Predicting misinformation and engagement in covid-19 twitter discourse in the first months of the outbreak. arXiv preprint arXiv:2012.02164

Widyantoro D, Wibisono Y (2014) Modeling credibility assessment and explanation for tweets based on sentiment analysis. J Theor Appl Inf Technol 70(3):540–548

Yamaguchi Y, Takahashi T, Amagasa T, Kitagawa H (2010) Turank: Twitter user ranking based on user-tweet graph analysis. In: International conference on web information systems engineering, Springer. pp 240–253

Zhang Y, Tiwari P, Song D, Mao X, Wang P, Li X, Pandey HM (2021) Learning interaction dynamics with an interactive lstm for conversational sentiment analysis. Neural Netw 133:40–56

Zubiaga A, Liakata M, Procter R (2017) Exploiting context for rumour detection in social media. In: International conference on social informatics, pp 109–123 Springer

Zubiaga A, Liakata M, Procter R, Bontcheva K, Tolmie P (2015) Crowdsourcing the annotation of rumourous conversations in social media. In: Proceedings of the 24th international conference on world wide web, pp 347–353