IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 27, 2021, accepted January 22, 2022, date of publication January 26, 2022, date of current version February 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3146712

HANN: Hybrid Attention Neural Network for Detecting Covid-19 Related Rumors

ABDULQADER M. ALMARS[1], MALIK ALMALIKI[1], TALAL H. NOOR[1], MAJED M. ALWATEER[1], AND ELSAYED ATLAM[1,2]

[1]College of Computer Science and Engineering, Taibah University, Yanbu, Medina 46411, Saudi Arabia
[2]Department of Computer Science, Faculty of Science, Tanta University, Tanta 31527, Egypt

Corresponding authors: Abdulqader M. Almars (amars@taibahu.edu.sa) and Elsayed Atlam (satlam@yahoo.com)

ABSTRACT In the age of social media, the spread of rumors is becoming easier due to the proliferation of communication and information dissemination platforms. Detecting rumors is a major problem with significant consequences for the economy, democracy, and public safety. Deep learning approaches were used to classify rumors and have yielded state-of-the-art results. Nevertheless, the majority of techniques do not attempt to explain why or how decisions are made. This paper introduces a hybrid attention neural network (HANN) to identify rumors from social media. The advantage of HANN is that it will allow the main user to capture the relative and important features between different classes as well as provide an explanation of the model's decisions. Two deep neural networks are included in the proposal: CNNs and Bidirectional Long Short Term Memory (Bi-LSTM) networks with attention modules. In this paper, the model is trained using a benchmark dataset containing 3612 distinct tweets crawled from Twitter including several types of rumors related to COVID-19. Each subset of data has a balanced label distribution with 1480 rumors tweets (46.87%) and 1677 non-rumors tweets (53.12%). The experimental results demonstrate that the new approach (HANN model) performs better results in terms of performance and accuracy (about 0.915%) than many contemporary models (AraBERT, MARBEART, PCNN, LSTM, LSTM-PCNN and Attention LSTM). Moreover, a number of software engineering features such as followers, friends, and registration age are used to enhance the model's accuracy.

INDEX TERMS Rumors, HANN, LSTM, attention Bi-LSTM, F-Score, accuracy, explainable rumors detection.

I. INTRODUCTION

In recent years, particularly social media has been proving to be a powerful tool for disseminating information at a faster rate than traditional networks. A rumor is an unconfirmed piece of text that spreads online and decreases trust in health authorities. Today's social media explosion has resulted in the spread of rumors that can threaten cyber security and social stability. During the Covid 19 pandemic, misinformation has spread rapidly through social networks and within communities [1]–[3]. The spread of COVID-19 has made individuals' not able to determine which information about the virus is trustworthy or not. Figure 1 shows a sample of rumors related to Covid 19 from ArCOV dataset [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

Discovering rumors is an interesting and significant problem. A variety of existing work is devoted to identifying rumors from social media. Existing techniques can generally be divided into two groups: (i) classical machine learning techniques [5]–[9], (ii) deep learning-based techniques [10]–[12]. Support Vector Machines (SVM) is one of the most popular rumor-detection algorithms. Chang *et al.* applied unsupervised clustering-based techniques to detect rumors tweets related to US elections [6]. Decision trees are also used to detect rumors [8]. Social media rumor detection using deep learning methods has been shown to be effective. Wu *et al.* [10] developed a sparse learning strategy for selecting discriminative characteristics and training the classifier for emerging rumors. Ma *et al.* [11] recently proposed recurrent neural networks for rumor classification, using sequential data in order specifically to identify the temporal and textual aspects of the spread of rumor,

which enables rumor detection to happen earlier and more accurately.

However, the main limitations of the current methods can be summarized in two points. First. when analyzing text for text classification, not every word in a sentence has the same importance. In other words, there are some words that are closely relevant to rumors, but others are irrelevant. The existing work considers that all words are significant when classifying texts as rumors or non-rumors. In the case of rumor detections, using attention techniques will help to know who generates a set of rumors and discover them easily. This mechanism is designed to ensure that important content receives more attention by assigning higher weights to certain keywords. Second, these deep learning approaches can classify or detect information to rumors or non-rumors without explaining why the model reached these decisions. Interpreting the reasons behind the model' decisions is critical and can help individuals understand why the model made such decisions. Third, the existing deep learning approach only utilizes textual information. Other software engineering features such as followers, friends, and registration age can help to detect rumors and enhance the model accuracy.

ID	Rumor in Arabic	Translated Text
1	إندونيسيا تسعين بالأشباح لإجبار الناس في البقاء في منازلهم في ظل جائحة كورنا.	Indonesia government uses ghosts to force people stay at home during Covid19 pandemic.
2	سبب انتشار فايروس كورنا يعود لانتشار شبكات الجيل الخامس.	5G networks are the reason for the spread of Covid19
3	الغرغرة بالملح وتنظيف الانف طريقة فعالة لحمايتك من فايروس كورنا.	Salted gargles and cleaning noise is an effective method of protecting yourself from Covid19.

FIGURE 1. Sample of arabic and English rumors in twitters.

To fill the gaps, this paper introduces a hybrid attention neural network (HANN) to identify and explain rumors detection from social media. The advantage of HANN is that it will allow the end-user to not only capture the relative and important features between different classes but also it provides an explanation of the model's decisions. The proposed model combines two deep neural networks with attention mechanisms: a Convolutional Neural Network (CNN) and a Bidirectional Long Short Term Memory (Bi-LSTM) network. The advantages of the proposed HANN model over the existing ones are summarized as follows:

- It provides a hybrid deep learning model to identify rumors from social media.
- It captures the relative and important features between different classes with an explanation of the model's decisions.
- The performance of the new approach achieved about 0.915% better than state-of-art approaches (AraBERT, MARBEART, PCNN, LSTM, LSTM-PCNN and Attention LSTM).
- A number of software engineering features such as followers, friends, and registration age were used to enhance the model's accuracy.

This study is organized as follows: Section 2 presents study literature. Section 3 and 4 discuss the system methodology. Section 5 presents data sets and system classification model and presents the evaluation results of the new model. Finally, section 6 focuses on the conclusion and points out future directions.

II. RELATED WORK

There is substantial interest in the detection of rumors in several fields, including data mining, machine learning, and natural language processing (NLP). This section re-views existing methods of detecting rumors in text content published on social media. We will focus primarily on detecting rumors from Twitter messages. In general, the current studies fall into two groups: (i) traditional machine-learning techniques, and (ii) deep learning techniques. We discuss, In this section, we discuss the literature work on rumors detection briefly.

A. MACHINE LEARNING METHODS

Traditional machine learning methods have been applied for rumour detection. A real-time approach to detecting rumors on Twitter was used by Suchita Jain *et al.* (2016) [13] by analyzing sentiment and semantics. They used verified News Channel accounts to classify rumors in real-time. According to Mao *et al.* [14], social media sentiment analysis was the most effective method for detecting rumors. To detect rumors, they combined shallow statistical features with deep statistical features and sentiment analysis. A rule-based heuristic method was proposed by Sivasangari *et al.* [15] which computed the sentiment polarities of each text. In order to distinguish a rumor from genuine content, VADER was used to determine the sentiment score value for the text.

Researchers also use Support Vector Machines and sentiment analysis to detect rumors [16]–[18]. A sentiment-based hybrid kernel SVM (SHSVM) classifier designed by Li *et al.* has been developed for detecting rumors [16]. A dictionary of emotions was used to analyze sentiments in comments on social networks. Based on the work by Zhang *et al.*, rumour detection was achieved by using shallow and implicit features [7]. They employed traditional machine learning approaches such as SVM (Support Vector Machine), Random Forest to classify tweets as rumors or non-rumors. Jin *et al.* [19] developed an approach for detecting the spread of rumors during 2016 U.S. elections. In addition, Word matching methods such as Word2Vec and Doc2Vec were used to identify tweets referring to two presidential candidates with verified rumor articles. In a study conducted by

Alqurashi *et al.* [20], a dataset of COVID-19 fake news that was spread on Twitter in Arabic was examined in detail. The methods examined in this study are the random forest classifier, the XGB, the naive Bayes, SGD, and the SVM. Alsudias and Rayson employed SVM, LR, and NB classifiers to distinguish rumor tweets from non-rumors [21], They applied their model to an Arabic dataset related to Covid 19. The best accuracy was achieved with LR using a count vector and SVM, which posted a result of 84.03%.

B. DEEP LEARNING METHODS

The use of deep learning methods has been proven to be effective on a variety of classification tasks. Unlike machine learning methods, deep learning approaches learn latent representations of input information to detect rumor from social media. Chen *et al.* [22] applied BERT model with TextCNN and TextRNN models to detect rumors. Training of the proposed model used data collected from 3737 rumors related to the COVD-19. According to the results, the proposed BERT model outperformed all other methods. Long Short-Term Memory (LSTM) is a popular algorithm for searching patterns in longer sequences [23]. LSTM holds the connection between various words in these sequences. Al-Sarem *et al.* [24] proposed a hybrid deep learning model. The proposed model uses a Long Short-Term Memory (LSTM) and Concatenated Parallel Convolutional Neural Networks (PCNN) to detect COVID-19-related rumors from Twitter data. Chen *et al.* [25] introduced a hybrid model for Cantonese rumors classification on Twitter. Bi-Directional Graph Convolutional Networks is proposed by Bain *et al.* to explore characteristics of rumors [26]. It leverages a GCN to learn the patterns of rumor propagation. a CNN+RNN model has been also introduced to detect fake Using user characteristics to create a feature vector, modeling five minutes of tweets into a time interval [27]–[30]. Other researchers combined convolutional neural networks (CNN) and long short-term memory (LSTM) to detect rumors based on the relationship between user and textual information [24]. As part of detecting rumors in the comments, CNN-LSTM models were used, and the sentiment was taken into consideration [31]. In the past few years, attention mechanisms have been incorporated into deep learning for classification purposes [10], [32], [33]. An experiment by Almars *et al.* demonstrated that application of attention mechanisms can improve depression detection from social media [33]. Yu *et al.* implemented an attention-based LSTM model to recognize and analyze speech emotions [32].

To sum up, previous researchers focused on classifying or detecting the information to rumors or non-rumors and could not explain rumors detection. To the best of our knowledge, this is the first research to utilize an attention mechanism address to detect rumours from Arabic content. Additionally, all studies do not take into account other factors like followers, friends, or age when detecting rumors in social media. To overcome these lacks, in this paper we suggest a hybrid attention neural network to identify and explain rumors detection from social media. Further enhancements are made to the model by leveraging software engineering features.

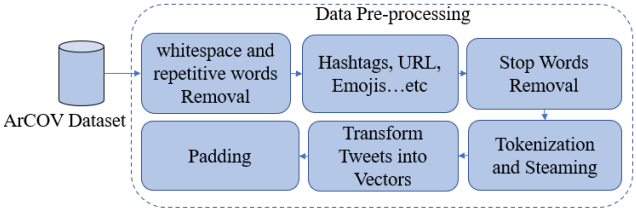


FIGURE 2. Data pre-processing steps.

III. METHODOLOGY

This paper introduces a hybrid attention neural network (HANN) to identify rumors from social media. The advantage of HANN is that it captures the relative and important features between different classes as well as provides an explanation of the model's decisions. The architecture of the proposed model is shown in Figure 3. The proposed model companies two deep neural networks: a Convolutional Neural Network (CNN) and a Bidirectional Long Short Term Memory (Bi-LSTM) network with an attention mechanism. The model extracts text features by using the CNN and then combines that with the Bi-LSTM and the attention mechanism for rumors detection. Below, we explain each step in detail.

A. INPUT LAYER

The pre-processing of data involves performing basic operations on the dataset before it is passed to HANN. In this step, raw data is transformed into an organized and useful representation that can be used for further analysis. As part of our model, the preprocessing step aims to eliminate noise and enhance rumor prediction. The main pre-processing consists of the following steps as shown in Figure 2:

- Eliminate whitespaces and repetitive words.
- Eliminate noises such as hashtags, URLs, replies, emojis, digits and punctuation.
- Eliminate stop-words.
- Apply segmentation, tokenization and steaming processes.
- Transform tweets into sequences of integers (vectors).
- Using zeros as padding to rescale the data.

B. EMBEDDING LAYER

In natural language processing, many feature extraction techniques have been proposed to determine the association and relationship between words. Bag of words and TF-IDF are statistical techniques used to determine the mathematical significance of words in documents. TF-IDF method was used in previous studies to analyse documents and find the significance of words in documents [18], [33], [34]. However, a number of studies used embedding techniques and showed

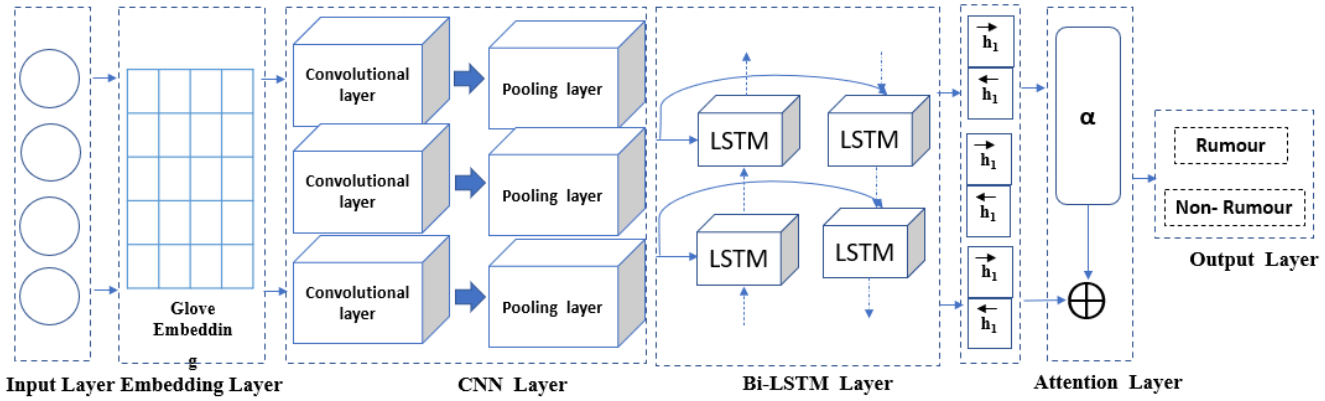


FIGURE 3. The architecture of the proposed model (HANN).

remarkable results. Embedding techniques can capture word associations and improve prediction accuracy [24], [26], [29]. Word embedding is one of the most popular techniques to represent text vocabulary. This technique can detect the context of a word within a document, as well as its semantic and syntactic similarity, and its relationship to other words. For learning to embed words from raw text, various models have been proposed, such as GloVe [35], Word2Vec [36] and FastText. The layer can be used to learn and store the embedding of words that can be used in the next layer. Based on the experiments, GloVe word embedding produced the best results. In this paper, we employed a pre-trained GloVe model for learning feature representations.

C. CNN LAYER

A CNN network is a type of deep learning network that produces excellent results and has been applied in several text classification takes. The goal of the CCN layer is to extract semantic features from the input text and reduce the number of dimensions. Standard CNN architecture layers consist of 3 convolutional layers, 3 pooling layers, and 1 fully connected layer. Each convolution layer is composed of a plurality of convolution kernels that convolve the input with pooling layers, and its calculation is shown in equation (1). Pooling layers are used to minimize the dimensionality of data and control over-fitting.

After applying different convolution layers, Several features are extracted from the data, but the extracted dimensions are very high. In order to reduce the feature dimension, a global max-pooling is applied at the end of each layer, and it produces the global best results from the entire network.

$$o_t = f(x_t * k_t + b_t) \tag{1}$$

where o_t is the output value after convolution, f is a RELU activation function which mathematically expressed as $f(x) = \max(0, x)$. x_t represents the input vector, k_t represents the weight of the convolution kernel, and b_t is the bias of the convolution kernel.

D. BI-LSTM LAYER

The Bi-LSTMs can process input sequences with variable length by two independent LSTMs (forward and backward). This section offers a quick overview of stander LSTMs and bidirectional LSTMs.

1) BASIC LSTM

Given input representations $x = \{x_1, x_2, x_3, \ldots x_t\}$, with length T, LSTM predicts the output sequence $y = \{y_1, y_2, y_3, \ldots y_t\}$, and compute the hidden state $h = \{h_1, h_2, h_3, \ldots h_t\}$, as follow:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$
$$c_t = f_t \odot c_{t-1} + t_i \odot tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$
$$h_t = o_t \odot tanh(c_t) \tag{6}$$

where h_{t-1} is the last hidden state, $x_t$is the current input, $W*$ are the weighted matrix and $b*$ represent the biases. σ is a sigmoid function and tanh is tangent function. $\odot$ refers to the element-wise product between two vectors.

2) BI-LSTMS

Basic LSTM is only able to remember contextual information based on previous information [37]. However, bidirectional LSTMs (Bi-LSTMs) [38] solve this problem by using the forward layer and backward layer to process the contextual information in both directions. The objective of bidirectional encoding is to encode rumor data in both directions. The output of the memory cell is calculated as follows:

$$\overrightarrow{h} = \{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_t}\} \tag{7}$$
$$\overleftarrow{h} = \{\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_t}\} \tag{8}$$
$$h_t = \overrightarrow{h}_t \oplus \overleftarrow{h}_t. \tag{9}$$

where $\overrightarrow{h}$ is the forward hidden state, $\overleftarrow{h}$ is the backward hidden state and $\oplus$ donates the concatenation operation. The output layer of the model is then updated by concatenating

$\overrightarrow{h}$ and $\overleftarrow{h}$. In this layer, the Adam optimizer was chosen, and the learning rate was set at 0.001.

E. ATTENTION LAYER

For the classification of text, understanding the relationship between the words is necessary so that the model can accurately label the text as ''rumors'' or ''non-rumors''. Words are obviously not equally important to a text representation; some features are more valuable than others. Instead of passing the hidden sequence to the next layer, we capture only relevant phrases using the attention mechanism. Employing attention mechanisms, we can interpret the significance of the different words, as well as the relative importance of the words themselves. Indeed, attention mechanisms emphasize important features by giving specific keywords a higher weighting. The attention weights of words are computed as follows:

$$A_t = \upsilon_w^T(W_w h_t + b_w), \tag{10}$$

$$A_t = softmax\frac{exp(A_t)}{\sum_t exp(A_t)}, \tag{11}$$

$$s = \sum_t a_t h_t. \tag{12}$$

where W_w is the weight matrices and b_w denotes the bias term, υ_w^T is a transposed weight vector. a_t is the normalized attention weights via softmax function, and a weighted sum of hidden representations is computed as s.

F. OUTPUT LAYER

The proposed model defines binary cross-entropy as a loss function for rumor detection. Afterwards, the attention layer's output is passed to the sigmoid layer with a value of either 0 or 1:

$$y = \begin{cases} 0 & p[0, 0.5] \\ 1 & Otherwise \end{cases} \tag{13}$$

where p is the predicted probability and y is the classification output where the predicted result with 0 representing non - rumor and 1 representing rumor.

IV. USERS' FEEDBACK TO OPTIMIZE HANN'S ACCURACY

Although a high accuracy of rumor detection can be achieved by HANN, room for higher accuracy is still possible through the use of users' explicit feedback (e.g., user's rating of posts as rumors or facts). The accuracy of machine learning systems can be significantly improved by working closely with users, which leads to a better understanding and trust from users [39]–[41]. Users can provide support to the HANN model by giving their explicit collective feedback on the model's classification accuracy and detection of rumors. This collective feedback will then be used to better train and tune the HANN's overall accuracy. For example, users could give points or ratings on the accuracy of HANN's classified social media posts. Furthermore, users can provide or receive feedback from other users on posts that the model failed to classify as rumors or facts. Users' ratings can be used to judge a post's quality, which will be used to educate HANN about the difference between rumors and facts. HANN's detection of rumors will be improved as a result, but this can also help minimize the negative effects of rumors on users and enhance the quality of information shared and produced on social media platforms.

It is, however, difficult to motivate users to give such feedback in a continuing manner, since the majority have little interest in doing so [42]. The concept of ''gamification'' is being used as a behaviour-change tactic to increase a user's motivation towards desired behaviours (for example, giving feedback about HANN's detection of rumors) [43], [44]. Among the common uses of gamification is taking the elements of a video game (e.g., points and levels) and applying them to a context that is not a game (e.g., an educational setting) [45].

Gamification has been successfully applied in several different environments, such as adopting a healthy lifestyle [46], Enhancing students' engagement with classes [47], increasing the quality and productivity in a business environment [48], etc. There are four common gamification elements in a non-game context [49]:

- Points: A number of gamification elements are based on points (e.g., levels and leader boards). The quality of a user's post (or a HANN's classified post) is deter-mined by ratings from other users in a social network. These gained or lost points will then be used to further train HANN to better distinguish rumor from facts. However, points should be used alongside other gamification elements to motivate users effectively [49].
- Digital Badges: Users can display these awards if they earn ''any kind of skill, knowledge, or achievement'' to show off their achievements [50] and have defined criteria [51]–[53]. Users may be able to collect digital badges, as an example, by accumulating a predetermined number of points based on the quality ratings, or by providing ratings on HANN's classification.
- Levels: In order to achieve levels, users need to earn points. Once they have earned a certain (predetermined) number of points, they can level up (i.e., unlock more software/game features) [54].
- Leader Boards: Users can create leader boards based on their achievements or points earned, or based on their progress towards a goal [52].

According to a recent study [55], there are several factors that influence how users perceive and respond to gamification elements in the context of detecting rumors on Twitter platforms, including:

- Privacy: The availability of a feature that allows users to rate each other's posts can be selected from a variety of privacy preferences.
- Notification: Users have various preferences on the style of notification they would receive one their posts are being rated or when a HANN's classified post needs to be rated. For example, some users would prefer to be

notified when only a negative rating is given on their posts.
- Gamification elements for online rumors: The majority of users would always prefer to have the option to use or deactivate such gamification elements (e.g., points) and not feel pressured by them.
- Social pressure: User relationships can negatively affect the objectivity of their ratings on posts when they are close to each other (e.g., a family member).

This sheds light on the need to carefully collect users' explicit and collective feedback that can be used to optimize HANN model in a manner that better suits users' preferences. By failing to do so, users' feedback on others' posts (or HANN's classified posts) could suffer, potentially leading to a failure of the whole usage of users' feedback to improve HANN's accuracy. To reach this, we adopt the application-independent conceptual framework proposed by [55] to gamify users' feedback collection process on the accuracy of HANN's classification and the detection of rumors that the model could fail to classify as rumors or facts (See Figure 4). The framework encapsulates the previously discussed differences of users' perceptions and needs towards the use of gamification elements to motivate them to give quality feedback on HANN's accuracy. This will guide software engineers on how to encourage users to provide their explicit and collective feedback to be used to further train HANN and potentially increase its detection accuracy of rumors.

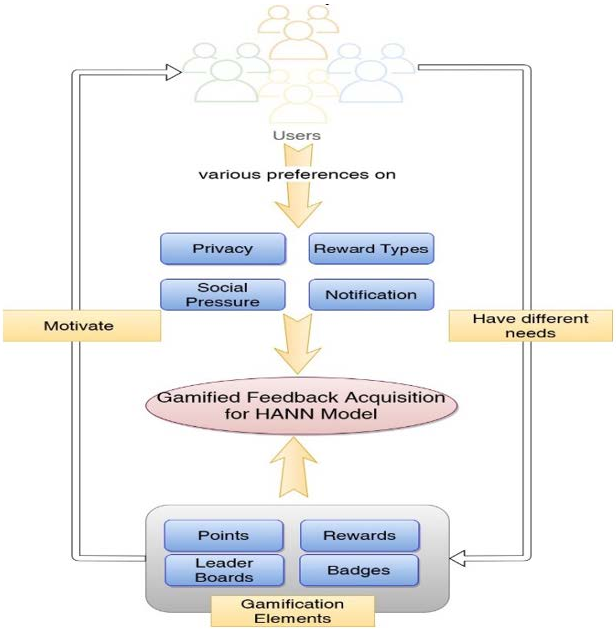


FIGURE 4. A conceptual framework for a gamified collection of users' feedbacks to optimize HANN model. Adapted from [55].

V. EXPERIMENTAL EVALUATION

This section examines our HANN model's performance using the ArCOV dataset [4]. We first describe the dataset used to evaluate our model. Based on the accuracy (Acc), sensitivity (Sen), specificity (Sep), and F1 score of our model, we measured its performance against other approaches.

A. DATASET DESCRIPTION

Experimental in this study uses benchmark dataset called ArCOV [4]. This dataset contains 3612 distinct tweets crawled from Twitter covering the period from 27 January to 30 April 2020. In order to search for tweets, the following keywords are used: (e.g., "Corona", "COVID-19", "the new Coronavirus", "the killing virus"). The dataset included several types of rumors related to COVID-19. Each subset of data has a balanced label distribution with 1480 rumors tweets (46.87%) and 1677 non-rumors tweets (53.12%) as shown in Figure 5. Certain preprocessing steps were conducted on the dataset. Noise is anything that decreases the effectiveness of the algorithm and prevents one from getting insights from the text. Stop-words, white-spaces, hashtags, and URLs are considered noisy text data. WordCloud modules have been utilized to visualize the majority of words that have been portrayed in each sentiment (Rumors and non-Rumors) (see Figure 6).

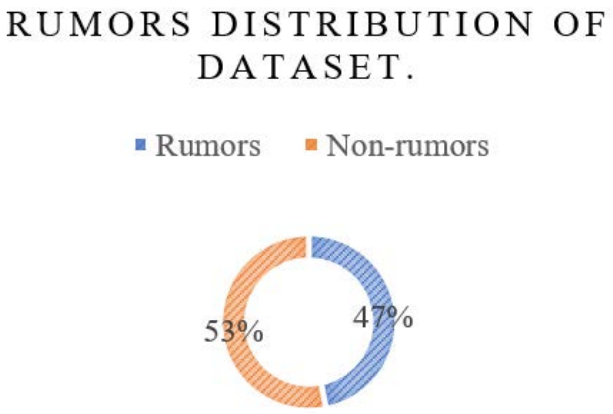


FIGURE 5. Label distribution for rumors and non-rumors tweets.

B. EVALUATION METHODS

To compare the efficiency and efficacy of various classification systems, a variety of measures can be used. The suggested model is evaluated using the following assessment metrics: accuracy (Acc), sensitivity (Sen), specificity (Sep), and F1 score. Using those measures, predictability can be computed by true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In our experiment, all methods are evaluated using 5 cross validations as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = \frac{TP}{TP + FP} \tag{15}$$

FIGURE 6. Most used words in rumors and non rumors tweets in dataset.

ID	Rumor in Arabic	Translated Text
1	كوريا الشمالية اعدام شخص يصاب بفيروس كورنا منعا لانتشاره بالبلد	North Korea killed a person to prevent the spread of coronavirus.
2	ينتقل فايروس كورنا عبر البعوض	Mosquitoes can spread the coronavirus
3	سبب انتشار فايروس كورنا يعود لانتشار شبكات الجيل الخامس	5G networks are the reason for the spread of Covid19
4	الغرغرة بالملح وتنظيف الانف طريقة فعالة لحمايتك من فايروس كورنا	Salted gargles and cleaning noise is an effective method of protecting yourself from Covid19.
5	فايروس كورنا يمكن معالجته بماء الثوم الطازج المغلي	Covid19 virus can be treated with boiling fresh garlic water

FIGURE 7. Visualized weight scores of different words.

$$F - Score = 2 \cdot \frac{precision * recall}{precision + recall} \tag{16}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

C. METHODS FOR COMPARISON

Our approach can mainly be compared with four typical models of rumor detection.

- AraBERT: A pre-trained BERT model that has been used recently for rumor detection from using a Arabic contents [56].
- MARBERT: is another pre-trained BERT model that has been used to analyze rumors from Twitter data. [57].
- PCNN: is another model that was commonly explored in previous rumors detection takes [24].
- LSTM-PCNN: This is a hybrid deep learning technique which analyzes texts from social media in order to detect rumors [26].
- LSTM: previous studies have utilized LSTM to perform rumor detection from social media and achieved good results [25].
- Attention–LSTM: This method utilizes attention mechanisms to find deep representations of data for efficient rumor identification [32], [33].
- The Proposed Model (HANN): A hybrid model that utilizes attention mechanisms to identify rumours from social media. Table 1 illustrates the description of HANN.

TABLE 1. Description of the proposed model.

Layer (Type)	Output Shape	Param #
Input	[(None,30)]	0
Embedding	(None,30,60)	635,760
bi-lstm-0	(None,30,60)	21,840
bi-lstm-1	(None,30,60)	21,840
Concatenate	(None,60)	0
Attention	(None,750)	124
Dense	(None,1)	61
Total Params		679,625
Trainable Params		679,625

D. RESULTS

To perform rumor detection, we employ attention techniques that extract key features from textual data. Using attention can help identify the importance of different words in our model as well as it provides an explanation of why the model reached a decision. To visualize the importance of different words in a sentence, we take the output of the attention layer and display the weights of the words with various colors. The color indicates how a specific word contributes to the final classification decision. Figure 7 shows a visual representation of the weights scores of different words. The darker colors indicate the most critical words. As seen from the visualization, the words that are highlighted in a sentence are specific to their correct class. Analyzing relative textual features can give some insight into rumor's classification decision. Hence, our model can reveal keywords that may be interpreted as rumor or non-rumor.

TABLE 2. Comparison of our proposed model with others.

Classifier	Embedding Type	Sensitivity	Specificity	F1-Score
AraBERT	-	0.721	0.717	0.713
MARBERT	-	0.754	0.735	0.739
PCNN	Word2Vec-CBOW	0.824	0.820	0.821
	Glove	0.861	0.857	0.857
LSTM-PCNN	Word2Vec-CBOW	0.854	0.851	0.850
	Glove	0.855	0.854	0.854
LSTM	Word2Vec-CBOW	0.897	0.839	0.839
	Glove	0.863	0.852	0.852
Attention-LSTM	Word2Vec-CBOW	0.889	0.889	0.882
	Glove	0.893	0.897	0.899
HANN	Word2Vec-CBOW	**0.901**	**0.907**	**0.906**
	Glove	**0.915**	**0.917**	**0.916**

TABLE 3. The accuracy, micro avg and Weighted avg of our proposed model compared with others.

Classifier	Embedding Type	Accuracy	Macro Avg	Weighted Avg
AraBERT	-	0.736	0.736	0.735
MARBERT	-	0.751	0.757	0.754
PCNN	Word2Vec- CBOW	0.844	0.846	0.844
	Glove	0.855	0.854	0.851
LSTM-PCNN	Word2Vec- CBOW	0.824	0.822	0.825
	Glove	0.5	0.844	0.844
LSTM	Word2Vec- CBOW	0.862	0.867	0.861
	Glove	0.846	0.849	0.843
Attention-LSTM	Word2Vec- CBOW	0.875	0.872	0.872
	Glove	0.893	0.891	0.898
HANN	Word2Vec- CBOW	**0.894**	**0.895**	**0.892**
	Glove	**0.915**	**0.915**	**0.917**

In this study, we have conducted a comparative study using state-of-the-art methods. A comparative study is carried out by comparing HANN with the following models: AraBERT, MARBEART, PCNN, LSTM, LSTM-PCNN, and Attention LSTM. Various metrics are used to evaluate the performance of each model such as: accuracy (Acc), sensitivity (Sen), specificity (Sep), F1 score, and ROC. Table 2 and 3 display the experimental results for our model compared with other traditional models. The values in Table 2 are listed in the order of lowest to highest accuracy values, with the highest values presenting last. It can be observed that the proposed method HANN achieves the greatest performance of sensitivity (Sen), specificity (Sep), and F1 score compared with the other approaches. Additionally, when the suggested model is compared to other approaches, the proposed model achieved the best results in terms of accuracy, micro average and weighted average as shown in Table 3. The suggested HANN model achieved 91.5%. Figure 8 shows the confusion matrix of HANN model. In addition, the AUC score shown in Figure 9 proves that HANN achieves good result which indicates the proposed model and the attention mechanisms are superior at extracting features, and this is why we use them in the our model to improve the accuracy of rumour detection. Further, HANN provides understandability on why a particular text should be flagged as rumor, compared to state-of-the-art methods.

E. DISCUSSION

Researchers have proposed several studies on rumor detection. However, previous studies focused only on classifying or detecting rumors or non-rumors, but couldn't explain how they detect rumours. In addition, none of the studies have taken other factors like the number of followers, friends, or even age into account when detecting rumors in social media. In this paper we suggest a hybrid attention neural network to identify and explain rumors detection from social media. As part of the attention mechanism, specific features that play a significant role in rumor detection are highlighted in order to increase the accuracy of predictions. Further improvements are made to the model by leveraging software engineering features. This sheds light on the need to carefully collect users' explicit and collective feedback that can be used to optimize HANN model in a manner that better suits users' preferences. By failing to do so, users' feedback on others' posts (or HANN's classified posts) could suffer, potentially leading to a failure of the whole usage of users' feedback to improve HANN's accuracy. To reach this, we adopt the application-independent conceptual framework proposed [55] to gamify users' feedback collection process on the accuracy of HANN's classification and the detection of rumors that the model could fail to classify as rumors or facts. In terms of performance and accuracy, the new approach (HANN model) outperformed many conventional

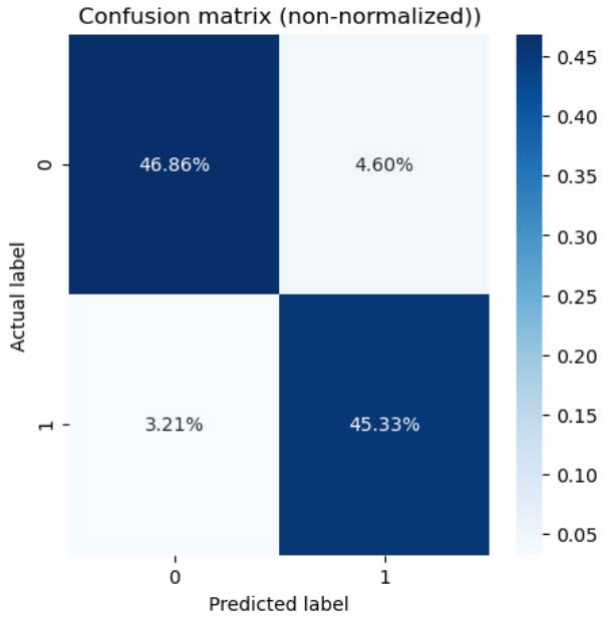


FIGURE 8. Confusion matrix arising out of the proposed model.

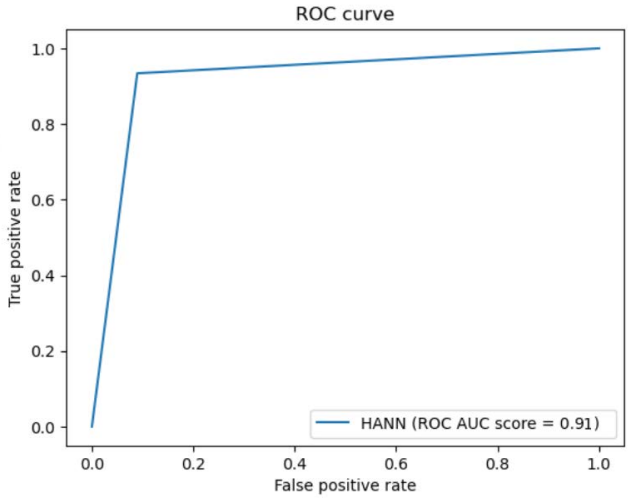


FIGURE 9. ROC AUC score for the proposed model.

approaches (LSTM, PCNN, LSTM-PCNN, and Attention Bi-LSTM). The suggested HANN model achieved (0.915%) accuracy as it's shown in Table 3. In summary, our proposed model's results provide two main findings. First, our model can successfully identify rumors from social media. A second advantage of HANN is that it can effectively highlight the significance of different features and provide explanations of classification results.

VI. CONCLUSION

Rumor detection is a crucial problem with far-reaching repercussions for the economy, democracy, and public health and safety. Applying traditional classification and deep learning algorithms to rumor identification cannot explain why and how texts are classified as rumor or non-rumor. This paper introduces a hybrid attention neural network (HANN) to identify rumors from social media. The advantage of HANN is that it provides an explanation of the model's decisions in addition to capturing the relative and important features between different classes. The proposed model includes two deep neural networks: CNNs and Bidirectional Long Short Term Memory (Bi-LSTM) networks with attention modules. According to experimental results, the new approach (HANN model) performed better than many contemporary models in terms of performance and accuracy (0.915%). The accuracy of the model was further enhanced by software engineering features such as followers, friends, and registration age. Future work could focus on predicting personality and society rumors with semantic structures. Furthermore, a software engineering method that guides software developers in implementing the adopted feedback acquisition framework will be proposed and tested on the HANN model.

REFERENCES

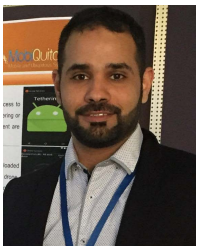
[1] M. S. Al-Zaman, "COVID-19-related fake news in social media," *medRxiv*, vol. 2, no. 1, pp. 100–114, 2020.
[2] S. Tasnim, M. M. Hossain, and H. Mazumder, "Impact of rumors and misinformation on COVID-19 in social media," *J. Preventive Med. Public Health*, vol. 53, no. 3, pp. 171–174, May 2020.
[3] A. Almars, X. Li, and X. Zhao, "Modelling user attitudes using hierarchical sentiment-topic model," *Data Knowl. Eng.*, vol. 119, pp. 139–149, Jan. 2019.
[4] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks," 2020, *arXiv:2004.05861*.
[5] A. Kumar and S. R. Sangwan, "Rumor detection using machine learning techniques on social media," in *Proc. Int. Conf. Innov. Comput. Commun.* Czech Republic: Springer, 2019, pp. 213–221.
[6] C. Chang, Y. Zhang, C. Szabo, and Q. Z. Sheng, "Extreme user and political rumor detection on Twitter," in *Proc. Int. Conf. Adv. Data Mining Appl.* Gold Coast, QLD, Australia: Springer, 2016, pp. 751–763.
[7] Y. Zhang, W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, "Detecting rumors on online social networks using multi-layer autoencoder," in *Proc. IEEE Technol. Eng. Manage. Conf. (TEMSCON)*, Jun. 2017, pp. 437–441.
[8] V. P. Sahana, A. R. Pias, R. Shastri, and S. Mandloi, "Automatic detection of rumoured tweets and finding its origin," in *Proc. Int. Conf. Comput. Netw. Commun. (CoCoNet)*, Dec. 2015, pp. 607–612.
[9] O. Araque and C. A. Iglesias, "An ensemble method for radicalization and hate speech detection online empowered by sentic computing," *Cogn. Comput.*, vol. 14, pp. 48–61, Feb. 2021.
[10] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 943–951.
[11] Z. Zhou, H. Zhang, and W. Pan, "Weibo rumor detection method based on user and content relationship," in *Artificial Intelligence in China*. Singapore: Springer, 2020, pp. 431–434.
[12] A. M. Almars, "Deepfakes detection techniques using deep learning: A survey," *J. Comput. Commun.*, vol. 9, no. 5, pp. 20–35, 2021.
[13] S. Jain, V. Sharma, and R. Kaushal, "Towards automated real-time detection of misinformation on Twitter," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 2015–2020.
[14] E. Mao, G. Chen, X. Liu, and B. Wang, "Research on detecting microblog rumors based on deep features and ensemble classifier," *Appl. Res. Comput.*, vol. 33, no. 11, pp. 3369–3373, 2016.
[15] V. Sivasangari, A. K. Mohan, K. Suthendran, and M. Sethumadhavan, "Isolating rumors using sentiment analysis," *J. Cyber Secur. Mobility*, vol. 7, pp. 181–200, Jan. 2018.
[16] L. Wy, "Research on microblog rumors detection pattern based on sentiment analysis," Chongqing Univ., Chongqing, China, 2016, p. 53.

[17] S. S. Roy, M. Biba, R. Kumar, R. Kumar, and P. Samui, "A new SVM method for recognizing polarity of sentiments in Twitter," in *Handbook of Research on Soft Computing and Nature-Inspired Algorithms*. Philadelphia, PA, USA: IGI Global, 2017, pp. 281–291.
[18] A. S. Alammary, "Arabic questions classification using modified TF-IDF," *IEEE Access*, vol. 9, pp. 95109–95122, 2021.
[19] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo, "Detection and analysis of 2016 US presidential election related rumors on Twitter," in *Proc. Int. Conf. Social Comput., Behav.-Cultural Modeling Predict. Behav. Represent. Modeling Simulation*. Washington, DC, USA: Springer, 2017, pp. 14–24.
[20] S. Alqurashi, B. Hamoui, A. Alashaikh, A. Alhindi, and E. Alanazi, "Eating garlic prevents COVID-19 infection: Detecting misinformation on the Arabic content of Twitter," 2021, *arXiv:2101.05626*.
[21] L. Alsudias and P. Rayson, "COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media?" in *Proc. 1st Workshop NLP COVID-19 ACL*, 2020, pp. 1–9.
[22] S. Chen, "Research on fine-grained classification of rumors in public crisis—Take the COVID-19 incident as an example," in *Proc. E3S Web Conf.*, vol. 179, 2020, p. 02027.
[23] A. Almars, X. Li, X. Zhao, I. A. Ibrahim, W. Yuan, and B. Li, "Structured sentiment analysis," in *Proc. Int. Conf. Adv. Data Mining Appl.* Singapore: Springer, 2017, pp. 695–707.
[24] M. Al-Sarem, A. Alsaeedi, F. Saeed, W. Boulila, and O. AmeerBakhsh, "A novel hybrid deep learning model for detecting COVID-19-related rumors on social media based on LSTM and concatenated parallel CNNs," *Appl. Sci.*, vol. 11, no. 17, p. 7940, Aug. 2021.
[25] X. Chen, L. Ke, Z. Lu, H. Su, and H. Wang, "A novel hybrid model for cantonese rumor detection on Twitter," *Appl. Sci.*, vol. 10, no. 20, p. 7093, Oct. 2020.
[26] A. Alsaeedi and M. Al-Sarem, "Detecting rumors on social media based on a CNN deep learning technique," *Arabian J. Sci. Eng.*, vol. 45, no. 12, pp. 10813–10844, Dec. 2020.
[27] S. Lv, H. Zhang, H. He, and B. Chen, "Microblog rumor detection based on comment sentiment and CNN-LSTM," in *Artificial Intelligence in China*. Singapore: Springer, 2020, pp. 148–156.
[28] F. Yang, S. K. Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, E. D. Ragan, S. Ji, and X. Hu, "XFake: Explainable fake news detector with visualizations," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 3600–3604.
[29] H. Karimi, P. Roy, S. Saba-Sadiya, and J. Tang, "Multi-source multi-class fake news detection," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1546–1557.
[30] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "DEFEND: Explainable fake news detection," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 395–405.
[31] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.
[32] Y. Yu and Y.-J. Kim, "Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database," *Electronics*, vol. 9, no. 5, p. 713, Apr. 2020.
[33] A. M. Almars, "Attention-based Bi-LSTM model for Arabic depression classification," *Comput., Mater. Continua*, vol. 71, no. 2, pp. 3091–3106, 2022. [Online]. Available: http://www.techscience.com/cmc/v71n2/45828
[34] K. Zhou, "Early rumour detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, 2019, pp. 1614–1623.
[35] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[36] Y. Goldberg and O. Levy, "word2vec explained: Deriving Mikolov et al.'s negative-sampling wordembedding method," 2014, *arXiv:1402.3722*.
[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[38] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[39] S. Stumpf, V. Rajaram, L. Li, M. Burnett, T. Dietterich, E. Sullivan, R. Drummond, and J. Herlocker, "Toward harnessing user feedback for machine learning," in *Proc. 12th Int. Conf. Intell. User Interfaces (IUI)*, 2007, pp. 82–91.
[40] T. Y. Lee, A. Smith, K. Seppi, N. Elmqvist, J. Boyd-Graber, and L. Findlater, "The human touch: How non-expert users perceive, interpret, and fix topic models," *Int. J. Hum.-Comput. Stud.*, vol. 105, pp. 28–42, Sep. 2017.
[41] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–15.
[42] M. Almaliki, N. Jiang, R. Ali, and F. Dalpiaz, "Gamified culture-aware feedback acquisition," in *Proc. IEEE/ACM 7th Int. Conf. Utility Cloud Comput.*, Dec. 2014, pp. 624–625.
[43] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining 'gamification,'" in *Proc. 15th Int. Academic MindTrek Conf. Envisioning Future Media Environments (MindTrek)*, 2011, pp. 9–15.
[44] P. Herzig, M. Ameling, and A. Schill, "A generic platform for enterprise gamification," in *Proc. Joint Work. IEEE/IFIP Conf. Softw. Archit. Eur. Conf. Softw. Archit.*, Aug. 2012, pp. 219–223.
[45] S. Nicholson, "A user-centered theoretical framework for meaningful gamification," in *Proc. Games+ Learn.+ Soc. 8.0*, Madison, WI, USA, vol. 1, 2012, pp. 223–230.
[46] D. Johnson, S. Deterding, K.-A. Kuhn, A. Staneva, S. Stoyanov, and L. Hides, "Gamification for health and wellbeing: A systematic review of the literature," *Internet Intervent.*, vol. 6, pp. 89–106, Nov. 2016.
[47] T. Pløhn and T. Aalberg, "Using gamification to motivate smoking cessation," in *Proc. Eur. Conf. Games Based Learn.*, 2015, p. 431.
[48] J. Simões, R. D. Redondo, and A. F. Vilas, "A social gamification framework for a K-6 learning platform," *Comput. Hum. Behav.*, vol. 29, no. 2, pp. 345–353, Mar. 2013.
[49] M. Lister, "Gamification: The effect on student motivation and performance at the post-secondary level," *Issues Trends Educ. Technol.*, vol. 3, no. 2, pp. 1–22, Dec. 2015.
[50] S. Abramovich, C. Schunn, and R. M. Higashi, "Are badges useful in education?: It depends upon the type of badge and expertise of learner," *Educ. Technol. Res. Develop.*, vol. 61, no. 2, pp. 217–232, Apr. 2013.
[51] J. Ahn, A. Pellicone, and B. S. Butler, "Open badges for education: What are the implications at the intersection of open systems and badging?" *Res. Learn. Technol.*, vol. 22, pp. 1–13, Aug. 2014.
[52] A. DomíNguez, J. Saenz-de-Navarrete, L. de-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes," *Comput. Educ.*, vol. 63, pp. 380–392, Apr. 2013.
[53] M. D. Hanus and J. Fox, "Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance," *Comput. Educ.*, vol. 80, pp. 152–161, Jan. 2015.
[54] G. Goehle, "Gamification and web-based homework," *Primus*, vol. 23, no. 3, pp. 234–246, Jan. 2013.
[55] M. Almaliki, "Misinformation-aware social media: A software engineering perspective," *IEEE Access*, vol. 7, pp. 182451–182458, 2019.
[56] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*.
[57] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," 2020, *arXiv:2101.01785*.

ABDULQADER M. ALMARS received the Ph.D. degree in computer science from The University of Queensland, in 2019. He is currently a Data Scientist. He is also working as an Assistant Professor and the Acting Head of the Information System Department, Taibah University, Yanbu. His current research interests include data mining, machine learning, data exploration and visualization, social network analytics, and health data analytics.

MALIK ALMALIKI received the B.Sc. degree in computer science from Taif University, Saudi Arabia (KSA), in 2008, the M.Sc. degree in advanced software engineering from Leicester University, U.K., in 2011, and the Ph.D. degree in software engineering from Bournemouth University, U.K., in 2015. He is currently an Assistant Professor in software engineering with the College of Science and Computer Engineering, Taibah University, Yanbu. His research interests include software engineering, adaptive software systems, and the engineering of social informatics, such as the systematic design of software-based solutions taking into consideration their interactions with related institutional and cultural contexts.

TALAL H. NOOR received the Ph.D. degree in computer science from the University of Adelaide, Australia. He is currently an Associate Professor and the Vice Dean of the Applied College, Badr Branch, Taibah University, Saudi Arabia. His research interests include services computing, security and privacy, trust management, social computing, and human–computer interaction.



MAJED M. ALWATEER received the Ph.D. degree in computer science from La Trobe University, Melbourne, Australia, in 2019. He is currently an Assistant Professor and the Acting Head of the Computer Science Department, Taibah University, Yanbu. His current research interests include pervasive computing, drone computing, and the IoT.

ELSAYED ATLAM received the Ph.D. degree in information science and intelligent systems from Tokushima University, Japan, in 2002. He is currently a Professor with the Department of Statistical and Computer Science, Tanta University, Egypt. His research interests include natural language processing, document processing, data mining and machine learning. He is a member of the Computer Algorithm Series of the IEEE and the Egyptian Mathematical Association. He was awarded by the Japan Society of the Promotion of Science (JSPS) Postdoctoral with the Department of Statistical and Computer Science, Tokushima University, from 2003 to 2005. He is an Editor Member of the *Information* journal of Tokyo and a reviewer for many international journals in his field.