IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 October 2022, accepted 24 October 2022, date of publication 26 October 2022, date of current version 7 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217522

RESEARCH ARTICLE

Arabic Rumor Detection Using Contextual Deep Bidirectional Language Modeling

NAELAH O. BAHURMUZ[1], GHADA A. AMOUDI[1], FATMAH A. BAOTHMAN[1], AMANI T. JAMAL[2], HANAN S. ALGHAMDI[1], AND AREEJ M. ALHOTHALI[2]

[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Ghada A. Amoudi (gaamoudi@kau.edu.sa)

This work was supported in part by the Deputyship for Research Innovation, Ministry of Education in Saudi Arabia; and in part by the King Abdulaziz University, DSR, Jeddah, Saudi Arabia, under Project IFPRC067-612-202.

ABSTRACT In today's world, news outlets have changed dramatically; newspapers are obsolete, and radio is no longer in the picture. People look for news online and on social media, such as Twitter and Facebook. Social media contributors share information and trending stories before verifying their truthfulness, thus, spreading rumors. Early identification of rumors from social media has attracted many researchers. However, a relatively smaller number of studies focused on other languages, such as Arabic. In this study, an Arabic rumor detection model is proposed. The model was built using transformer-based deep learning architecture. According to the literature, transformers are neural networks with outstanding performance in natural language processing tasks. Two transformers-based models, AraBERT and MARBERT, were employed, tested, and evaluated using three recently developed Arabic datasets. These models are extensions to the BERT, Bidirectional Encoder Representations from Transformers, a deep learning model that uses transformer architecture to learn the text representations and leverages the attention mechanism. We have also mitigated the challenges introduced by the imbalanced training datasets by employing two sampling techniques. The experimental results of our proposed approaches achieved a maximum accuracy of 0.97. This result demonstrated the effectiveness of the proposed method and outperformed other existing Arabic rumor detection methods.

INDEX TERMS Classification, deep learning, fake news, imbalanced data, machine learning, natural language processing, Twitter.

I. INTRODUCTION

Recently, as user-generated content is becoming popular in society, the possibility of spreading fake news and rumor stories increased. Many definitions of rumors have been found in the literature. One of the most used definitions found in most dictionaries is the one that considers rumors as ''a story or a statement whose truth value is unverified'' [1]. The definition states that rumors can be later deemed false or true; they do not have to be always wrong. Thus, the main feature of a rumor is that its truth value is uncertain and unverified at the publishing time [2]. Online social media platforms have become a significant source of news and up-to-date information, as most people acquire news from these platforms rather than traditional media channels. The importance of these platforms arises from the ubiquity and flexibility that allows anyone to instantly post and share information with a large group of audiences and allows them to gather information. This open nature of social media platforms and the unrestricted way for users to share information encourage information to spread quickly across the social network regardless of its validity or credibility. In addition, it gives fertile ground for rumormongers to share and spread rumors, which causes a high-velocity flow of rumors that grow unexpectedly and spread rapidly [3].

Readers on these platforms are dealing with a large amount of new information every moment and may be unable to verify its validity. One of the most used platforms in the

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

Arab region is Twitter which allows its users to publish limited characters messages called ''tweets''. This platform has become an ideal place for misinformation, fake news, and rumor propagation [4]. For instance, in 2013, there was a tweet posted on Twitter by the Associated Press (AP) official Twitter account, which was hacked at that time, stating that ''Explosion at White House with two bombs and the President of the USA being injured in the attack''. In just six minutes, the tweet spread quickly to millions of people, which resulted in a dramatic, even so brief, crash in the stock market. Although it was debunked soon, it harmed people and society, which shows the danger of spreading such rumors on social media platforms [5].

For Arabic content, there is a contribution from an independent Saudi Arabian project developed in 2012 known as The No Rumors Commission [6]. The main goal of this project is to counter rumors and sedition that are trending on different social media platforms. Also, it exposes the publishers of these rumors and clarifies the truth with the sources and evidence [6]. Their method is to publish a post on their verified Twitter account containing the trending rumor with the original authentic content. Generally, although their approach has gained wide popularity among users in social media and society, they depend only on the manual effort of their team and rely on human observation of trending rumors.

Additionally, not all social media users in Saudi Arabia follow the account of the No Rumor Commission or even know that it exists. Thus, there is a need for an automated system that assists social media users in checking the validity of news and information smoothly and effectively. Automatic rumor detection has attracted researchers' attention and has become one of social media analytics' most active research areas [4]. Many studies have contributed to this domain using several methods and approaches. Recently, researchers have adopted artificial intelligence techniques to develop detecting systems. Modern techniques use the deep learning subfield of artificial intelligence that focuses on creating deep neural network models capable of making accurate data-driven decisions [7]. Deep learning models train the computer to generate results using existing examples. This feature enables experimenting with deep models, especially when large and complex datasets are available for analysis [8].

Rumor detection approaches fall into three types based on the data [9]. The first method is content-based, focusing on the post's textual content and related user comments [10]. The second method is feature-based that utilizes several non-linguistic features such as profile information and the number of tweets [11]. The last one is propagation-based methods which use patterns in tweet propagation to identify rumors [12]. Detecting rumors in English language posts on Twitter is an active research area. However, for Arabic content, few studies contributed to rumor detection. This research study aims to build a framework to classify Arabic news to find rumors from social media posts. The objectives of the study are the following:

1. To optimize the rumor detection task by applying two recently developed transformer-based models.
2. To consolidate the findings by testing the model on multiple datasets and articulating the results.

Our focus is on Arabic tweets that are posted on Twitter platforms. To evaluate the framework, we used Twitter data related to the coronavirus COVID-19. This study is organized as follows: section two explains the background and basic terms; section three explains the related work; section four presents the methodology; section five illustrates the results; section six shows the discussion; and finally, section seven concludes the study and highlights future research directions.

II. BACKGROUND

A. DEEP LEARNING

Deep learning (DL) is a branch of Machine learning (ML), where the latter is a class of artificial intelligence that constructs a mathematical model based on sample data, called ''training data'', to make decisions or make predictions without being explicitly programmed to perform the task [13]. ML primary goal is to create theories and procedures, learning algorithms, that allow machines to learn and adapt from experience [14]. ML has three learning styles: supervised, unsupervised, and reinforcement learning [15].

DL is a machine learning approach that simulates the human brain's functioning and how they are structured. DL models are built based on an artificial neural network to learn from data and predict outcomes of unseen samples. The neural network framework consists of three types of layers: the input, output, and hidden layers [8]. Among the most substantial aspects of neural networks is the ability to obtain each layer's model parameters according to the training data [13].

One of the most popular models in DL applications is Convolutional Neural Networks (CNN). CNN works by shaping input data into a two-dimensional matrix, like time series or image pixels. Another standard layer in DL is the long short-term memory (LSTM), which learns long-term dependencies in sequential data such as text and time series [16].

B. TRANSFORMER-BASED TRANSFER LEARNING

Recently, many researchers have developed pre-trained models that effectively improve prediction performance on many Natural Language Processing (NLP) tasks [17]. These models are called transformers and are based on neural networks trained on large textual data using unsupervised objectives [18]. Transformers include stacked blocks of encoders and decoders and employ the self-attention mechanism [19]. The attention mechanism has been widely adopted in several deep learning models, especially for transformer models. It works by directing the models to focus selectively on specific and relevant information and ignoring other irrelevant information [20]. Some researchers attributed the great success of transformer-based models to the attention element that allows contextual information to be captured by the network through the whole sequence. A multi-head attention layer

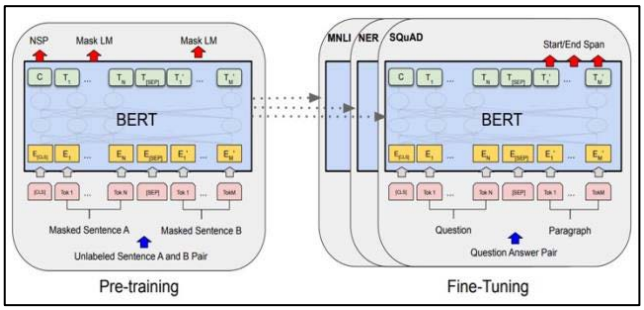


FIGURE 1. Architecture of pretraining and Fine-Tuning BERT model [25].

contains various scaled dot attention layers. It runs through an attention mechanism several times in parallel, and then attention layers are concatenated at the end of the process [21]. Applying a pre-trained language model to another classification task is called transfer learning. The most common transfer learning techniques are feature-based and fine-tuning [22]. The features-based approach uses task-specific architectures that include pretraining real-valued embedding vectors in different levels (word, sentence, or paragraph). The embeddings are then fed to a specific model as additional features [23]. While the second approach, fine-tuning, involves minimal task-specific parameters and is done by tweaking all the pre-trained parameters to be trained on the required task [22]. The latest work confirms that fine-tuning often performs better than feature-based transfer on text classification tasks [24]. The BERT model is one of these pre-trained models that improve the fine-tuning-based approaches [22].

BERT, Bidirectional Encoder Representations from Transformers, differs from most language representation models in that it is designed to pre-train deep bidirectional representations from an unlabeled text using the join condition on both the right and left contexts in all layers. It is focused on learning the context of word embeddings. The model was trained with multi-task objectives: the masked language modeling task that enables the representation to fuse the right and left contexts and the next-sentence prediction task that jointly pretrains text-pair representation. The fine-tuning approach is straightforward in BERT since the self-attention mechanism allows it to model several classification tasks by swapping suitable inputs and outputs. BERT model has trained over 3.3 billion words of English corpus [22]. BERT utilizes 12 layers of encoder networks, 768 hidden state dimensions, 12 attention heads, and 512 maximum sentence lengths. Figure 1 shows the architecture of BERT during pretraining and fine-tuning operations.

Language models (LMs) can be multilingual or single, depending on the corpus used to pretrain the model. Even though multilingual models were introduced to serve many languages, these models have some limitations; one of these is the costly inference process due to the enormous size and variety of non-English data involved in the pretraining stage [26]. Several experiments showed that the multilingual models perform well with many languages; however, recent studies showed that single-language models perform significantly better since they are pre-trained on a large corpus of specific languages resulting in better language understanding.

1) ARABERT MODEL

In pursuing the same success that the BERT model did for the English language, authors in [25] developed the AraBERT model by pre-training the BERT model on a large-scale Arabic corpus from Wikipedia containing 70 million sentences and 3 billion words. They evaluated the model on three NLP downstream tasks: sentiment analysis, question answering, and named entity recognition. The experiments on these Arabic NLP tasks showed that the AraBERT model achieved superior performance on most tested tasks compared to various baselines, including previous single-language and multilingual approaches [25].

2) MARBERT MODEL

MARBERT is an Arabic-focused Transformer LMs developed in 2021. It is pre-trained on massive and different datasets (1 billion Arabic tweets) to facilitate transfer learning on Modern Standard Arabic (MSA) and Arabic dialects. MARBERT uses the same network architecture as the BERT model but excludes the next sentence prediction objective due to the word count limit in tweets. MARBERT is trained using data from the Twitter platform, which includes both MSA and diverse Arabic dialects, whereas AraBERT is only trained on MSA data. MARBERT was evaluated on six NLP tasks, sentiment analysis, topic classification, dialect identification, question answering, named entity recognition, and social meaning. The results of these six tasks showed that MARBERT is significantly better than AraBERT, according to [26].

C. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a set of computational techniques that automate the analysis and representation of human languages, motivated by theory. NLP uses deep learning to understand human speech to perform several practical tasks. NLP focuses on natural language translation, text summarization, information extraction, topic modeling, information retrieval, text classification, and opinion mining [15]. In the past two decades, and due to the wide use of the World Wide Web, NLP has been utilized in many applications of real-world problems [27]. Text classification is one of the NLP tasks that classify a given text into some predefined classes. Some real-world text classifications are fraud, bot, spam detection, emergency response, and classification of commercial documents [24].

1) ARABIC NATURAL LANGUAGE PROCESSING

Arabic is an essential language as it is the native language of Arabic countries, with around 414 million people [28]. There are three primary forms of Arabic: Classical Arabic (CA), Modern Standard Arabic (MSA), and informal Arabic or dialects. The oldest form is the CA, the language of traditional

Arabic books and the Holy Quran. MSA is used for formal writing and communication in most Arabic countries [29]. The most used form of Arabic is informal Arabic, which includes variations depending on the country and sometimes on the region [30]. Various Arabic language studies were conducted to develop several Arabic NLP (ANLP) applications and automatically analyze text in multiple domains [31].

III. RELATED WORK

A. RUMOR DETECTION

Many studies proposed solutions and frameworks to detect rumors and spam content using various ML and DL approaches. Most rumor detection studies have used several ML algorithms that employ statistical features or feature engineering. These algorithms work by applying several machine learning-based classifiers after extracting multiple features from textual data [32].

In the health domain, authors in [33] developed a model that detects rumors from Twitter data. The approach is focused on each post and extracts features such as influence potential, network characteristics, and personal interest. Then, six different classifiers were employed, and the finding showed the framework could correctly detect 90% of rumors. Another approach presented by Buntain and Golbeck [34] automatically classifies popular threads on Twitter as accurate or inaccurate. The authors developed a method for fake news detection by learning to predict accuracy assessments. They trained the model on two public Twitter datasets. The results showed that their model correctly classifies two-thirds of the fake news stories posted on Twitter and outperforms previous work in the same domain [34]. Authors in [35] presented a hybrid neural network model that considers three aspects of rumor detection on Twitter: users, contents, and propagation. They used graph convolutional networks to model users' behaviors. Then, they used a recursive neural network to build a propagation tree encoder that represents the propagation tree of the contents. The last module is the integrator which combines the output of the above modules to recognize rumors.

All previous studies applied different ML techniques to the dataset to detect rumor content on social media platforms. However, feature-based and feature engineering methods have some limitations, requiring tedious manual efforts. Furthermore, the models trained by certain hand-crafted features have difficulty performing well under various rumor detection scenarios. Because of that, some researchers have been attracted to utilizing deep neural networks or deep learning to develop end-to-end models that automate classification tasks such as rumor detection [32].

Many studies employed DL techniques to handle raw text in different detection domains, such as classifying content on social media platforms. Authors in [2] combined DL and representation learning algorithms to automatically recognize emerging rumors during breaking news diffusion on the Twitter platform. Their approach flagged tweets with unverified information regardless of their truth value, based only on a tweet's text. They used semi-supervised learning and built an LSTM-RNN-based model. Another approach proposed by [36] combines joint text and propagation structure representation learning to develop a rumor detection framework. The process applied in this work took advantage of the texts in original tweets and the propagation structures of all tweets. Next, a CNN-based model was built to capture tweets' textual and propagation features. The result achieved an accuracy of 0.80 and 0.85 on two datasets and 0.95 on a third. Veyseh et al. [37] presented a semantic graph approach that used DL techniques to detect rumor news based on modeling the semantic relations between the posts and their replies. The model works by learning the implicit associations through the primary tweet and its replies based on their content. The authors compared their proposed model to deep learning and feature-based models. The results showed that deep learning models outperformed feature-based models for rumor detection. In addition, they showed that integrating implicit semantic relations through tweets in a thread achieved outstanding performance.

A recent line of research focused on pretrained and transformer-based models for various classification tasks, as these models showed an outstanding performance compared to traditional approaches [17]. Several studies demonstrated that pre-trained models on the vast corpus could learn universal language representations, which are beneficial for downstream NLP tasks and eliminate the need to train a new model for every new task [24]. Wani et al. [38] presented another DL approach that utilized BERT, CNN, and LSTM. They evaluated these supervised text classification algorithms on a COVID-19 fake news dataset. In addition, they assessed the significance of the pretrained language model and distributed word representations employing an unlabeled corpus of COVID-19 tweets. The best accuracy achieved was 0.98. Authors in [9] developed a transfer learning model where a CNN-bidirectional-LSTM (CNN- Bi-LSTM)- and BERT-based deep neural models are used to detect rumors early based on the tweets and their comments by other users. Results showed that the BERT-based model is more effective and outperforms start-of-the-art rumor and detection models. Another study [39] proposed an approach of fine-tuning (RoBERTa and CT-BERT) for the fake news detection task. The strategy aims to address the problems related to transformers models and their inability to make an actual distinction between whether the news is real or fake. The first stages concerned the token vocabulary and distinguishing the hard-mining samples, which are typical for fake news. Then, to improve the model's robustness, they involved adversarial training. Finally, they extracted the predicted features by two BERT models: RoBERTa, the universal language model, and CT-BERT, the domain-specific model. They integrated them with one multiple-layer perception to integrate fine-grained and high-level specific representations. The results demonstrated superior performances compared to other methods, and the best weighted average F1 score was 0.99. Similarly, authors in [19] used eight different pre-trained models named

GPT-2, XLNet, BERT, RoBERTa, DistilRoBERTa, ALBERT, Bart, and DeBERTa. They fine-tuned them by adding additional layers to build a stacking ensemble classifier. The best model achieved 0.98 accuracy and a 0.98 F1 score.

B. ARABIC TEXT CLASSIFICATION

A long line of research on ANLP was found in the literature in different fields, ranging from text classification to encryption and decryption methods [40]. Several ML approaches were applied to tasks on social media platforms. Text classification is the process of assigning a category or class to the text from a set of predefined classes according to a specific context [41]. Generally, researchers applied the traditional text mining pipeline and then used various classification models. Many classifiers are used, such as Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR). Alzanin et al. [42] proposed a method to classify Arabic tweets based on their content and linguistic characteristics. Five classes were considered, question, news, wish, conversation, and others. The study then applied two textual representations, Word2vec word embedding and TF-IDF, and three classifiers, SVM. Gaussian NB and RF. The study achieved an F1 score of 0.98. In [43], the authors presented a method for detecting Arabic text mentioning crimes using classification algorithms such as decision trees, SVM, complement NB, and K-nearest neighbors. Additionally, they evaluated different techniques of feature extraction, including n-gram, light stemming, and root-based stemming. Their results showed that the highest accuracy was 0.91, gained by applying the SVM with trigram.

Some studies adopted DL approaches, such as [44]. The study proposed a multi-label classification framework for Arabic text on Twitter. Two DL methods were used, CNN and Recurrent Neural Networks (RNN), and they achieved 0.90 accuracy scores. Another work used CNN for racism detection in Arabic tweets [45]. The study's findings showed that using CNN-based deep learning models is suitable for various large datasets and binary classification of a specific task. Alharthi et al. [16] presented a real-time model to identify low-quality tweets and accounts that make these contents on Twitter. They extracted the Arabic dataset from Twitter using Twitter API and then applied both CNN and LSTM techniques. Their results showed superior performance in detecting low-quality tweets compared with other real-time systems.

In the sentiment analysis domain, authors in [46] proposed an approach using the transformer-based BERT model, which integrates an Arabic BERT tokenizer instead of a basic BERT tokenizer. They tested the technique using five public datasets; the best accuracy was 0.96.

C. ARABIC RUMOR DETECTION

Several studies targeted Arabic rumor detection and control on social media platforms. Floos et al. [47] focused on Arabic fake tweets and used ML and text representation. The dataset was divided into two separate files: rumors and news. Then, TF-IDF (Term Frequency – Inverse Document Frequency) was applied to determine the terms in the documents. The finding showed an accuracy of 0.6 in detecting rumors. Another proposed system [48] used both semi-supervised and unsupervised learning to filter Arabic rumors on Twitter. They adopted a semi-supervised learning approach to reduce the human labor needed for labeling and to avoid bias. The proposed model achieved an F1 score of 0.78. In [49], the authors presented a supervised ML model to detect Arabic news articles based on their contexts. They used textual features, including linguistics, polarity, emotion, and part of speech. The model achieved 0.86 prediction accuracy and outperformed humans in the same task.

Another study applied DL techniques to detect Arabic rumors, which sometimes perform better than traditional ML models [50]. In [51], the authors presented approaches for detecting misinformation on Arabic Content on Twitter. They examined several machine learning and deep learning classifiers to recognize Arabic misinformation related to COVID-19 automatically using an annotated Arabic dataset and employed different features, including word embeddings (word2vec and FASTTEXT) and word frequency. The results showed that using word embeddings increased the performance of the classifier. In addition, optimizing the area under the curve (AUC) improves the performance of the models. Reference [31] investigated the use of standard ML and DL models to detect Arabic rumors. The study compared seven optimizers in the DL experiments. Their results stated that using classical ML without the stacking technique enhances the effects of predicting the rumors. In addition, the best results were obtained using both LSTM and Bi-LSTM with the root mean square propagation optimizer. Sorour and Abdelkader [52] proposed a method to detect Arabic news based on a hybrid DL (CNN-LSTM) model. They used a publicly available dataset called ANS. Their approach achieved a 0.81 accuracy score.

Authors in [53] studied neural networks and transformer-based models, such as AraBERT, MARBERT, ArElectra, Arbert, and QARiB. For DL models, they applied CNN, RNN, and gated recurrent unit (GRU). The experiment result revealed that QARIB scored the best F1 score of 0.95. In addition, it proves that transformer-based models outperform neural network-based models. Reference [54] an approach to detect rumors in machine-generated rumors was proposed. Four transformer-based models were employed: AraBERT, mBERT, XLM-RLarg, and XLM-RBase. They generate fake news by replacing one or two words from true news stories, employing a word embedding model. Then, classification models were applied to the newly developed dataset. The results showed the best F1 score of 0.70 using the XLM-RLarg classifier. A summary of Arabic Rumor detection literature is presented in Table 1.

As presented in the literature review above, many attempts have been made to detect and combat rumors in English and Arabic social media posts using ML, DL, and transformer-based learning. In this work, an evaluation of

TABLE 1. Summary of arabic rumor detection literature.

Study	Dataset	Approach	Result
[47]	New Arabic dataset from Twitter	ML, TF-IDF	Accuracy 0.67
[48]	Arabic rumor-non-rumor tweets	ML, Semi-supervised expectation–maximization	Accuracy 0.786
[49]	Arabic fake news articles	supervised ML	Accuracy 0.86
[51]	COVID-19 misinformation	DL, eight different traditional and DL models	Accuracy 0.86
[31]	ArCOV19-Rumors	ML and DL models	Accuracy 0.80
[52]	ANS	DL	Accuracy 0.81
[53]	ArCOV19-Rumors, COVID-19- Fakes, AraNews, ANS corpus	A comparative study, transformer-based models, and neural networks	F1 score 0.95
[54]	AraNews	Transformer-based models	F1 score 0.70

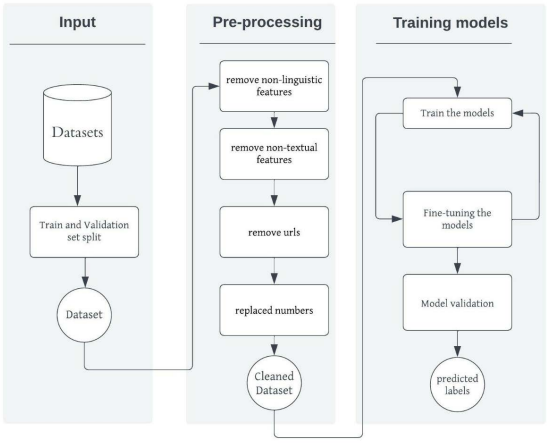


FIGURE 2. Proposed framework.

two transformer-based models was conducted on three different Arabic rumor datasets. We also dealt with the data imbalance issue to provide a more accurate and reliable result.

IV. METHODOLOGY

This research aims to develop a rumor detection framework by classifying Arabic tweets. The proposed model is a content-based model that focuses on the textual contents of Twitter's posts. The input to this proposed model is a stream of Arabic tweets, and the output is the class of the post, rumor or not rumor, as shown in Figure 2. This section outlines the steps to create the framework to achieve the study's objectives.

A. DATASET

This study used three Arabic public datasets to train the proposed models. The first dataset is named COVID-19 misinformation [51], which contains tweets extracted from Twitter for the four months of January 1, 2020, to April 30, 2020. The authors provided a list of the most common Arabic keywords (30 words) related to COVID-19 and used them to extract the tweets, including coronavirus, outbreak, pandemic, home quarantine, and social distancing. Then, they obtained relevant Arabic tweets by filtering the Twitter stream based on the selected keywords. To label the tweets, native Arabic speakers were hired and were informed about World Health Organization's guidelines regarding COVID-19 disease cures and measures. The dataset contains 8,783 tweets, 7,475 are original news, and 1,311 are fake news.

The second dataset provided by [55] is called Fake News Detection, a manually annotated dataset from Twitter's platform. The authors collected tweets related to COVID-19 by defining a list of related hashtags, from January 1, 2020, until May 31, 2020. Then they filtered these tweets to keep only tweets relevant to the pandemic and containing fake news keywords. Three annotators were involved; two completed the annotation, while the third evaluated the labeling and resolved the conflicts. The dataset includes 1,537 tweets (835 fake and 702 genuine).

The third dataset, Arabic rumor-non-rumor tweets, was developed by [48]. A total of 271,000 tweets were gathered, consisting of 88 events of non-rumor and 89 events of rumor. The rumor topics were obtained from anti-rumors authority [6] and a popular daily newspaper, Ar-Riyadh (http://www.alriyadh.com/).

We performed an exploratory data analysis for all datasets using Python to understand and prepare the datasets. First, we preserved only two columns, the Tweet text and the label, and removed the other columns. The values of the second column were standardized to contain binary values 0 and 1, in which 0 presents rumor content and 0 presents real news. The second step was removing redundant tweets. The third dataset contains 271,000 tweets and 78 columns (features). After removing the redundant tweets, 36,308 tweets were left. The random shuffle was applied to train the model, and data was split into an 80%-20% ratio for the train and test sets, respectively.

B. DATA PREPROCESSING

In models that deal with natural language processing classification tasks, selecting the appropriate text preprocessing methodology is essential and critical. The preprocessing methods, such as correcting misspelled words, removing duplicated letters and words, or normalizing text, may increase the accuracy of the classification tasks, or in some cases, it might increase the complexity of model computational and processing time [51]. Accordingly, we have considered many main preprocessing techniques in the data cleaning phase in our proposed system. To prepare the data for the training phase, we eliminated the following:

TABLE 2. Sample of arabic tweets with english translation.

Tweet	Dataset	Label
Arabic: " الشمس الحارقة تقضي على فايروس كورونا"	COVID-19 misinformation dataset	1
English: The hot sun kills Corona virus		
Arabic: " علاج كورونا ببلازما الدم من المتعافين قيد الدراسة في المملكة"		0
English: Corona treatment with blood plasma from recovered patients under study in the Kingdom		
Arabic: " دراسة صينية اثبتت ان فيروس كورونا لا يصيب اصحاب البشرة السودا"	Fake news detection	1
English: A Chinese study has proven that the Corona virus does not affect people with black skin		
Arabic: " مجففات الايدي غير فعالة في القضاء علي فيروس كورونا"		0
English: Hand dryers are ineffective in eliminating the coronavirus		
Arabic: " شجاعة شاب سعودي ينقذ طفلا في بلجيكا قبل أن يدهسه القطار"	Arabic rumor-non-rumor tweets dataset	1
English: The bravery of a Saudi young man saves a child in Belgium before being run over by a train		
Arabic: " عاجل اكتشاف أكبر حقل للنفط في تاريخ البحرين"		0
English: Urgent discovery of the largest oil field in the history of Bahrain		

a) Non-linguistic features, including user mentions starting with @, retweet signs, punctuation, and hashtags.
b) Non-textual features such as images, videos, and emojis.
c) URLs.
d) Numbers, were replaced with representative words.

Data cleaning was achieved using over-the-shelf tools, including CAMeL [56] and Farasa [57]. CAMeL is a Python library for Arabic natural language processing. It provides utilities for preprocessing, dialect identification, morphological modeling, named entity recognition, and sentiment analysis. Farasa is an Arabic word segmenter used in the segmentation and stemming steps.

C. EXPERIMENTAL SETUP

The implementation of the rumor detection model is performed using Python language as it supports a large variety of APIs that would facilitate the workflow. PyTorch was chosen as the deep learning framework. In this study, we focus on fine-tuning two of the most recent pre-trained LMs for the Arabic language, AraBERT, and MARBERT, to develop the rumor detection model. These pre-trained models were chosen due to their competitive performance [25], [26]. We conducted several hyperparameter optimization experiments on both AraBERT and MARBERT models to achieve the best performance with these deep models. AraBERT uses the base version of the BERT model. AraBERT's objective function aims to learn each word's context by predicting a word in any sentence. It is trained in an unsupervised manner, in which a few words from text sentences are kept hidden, and the model is forced to predict them. This technique helps in understanding context words in each sentence. AraBERT masked 15% of the words from the entire data in the MLM task. 80% of words of those selected words were replaced by the [MASK] token, 10% of those words were replaced with random tokens, and the last 10% were replaced with original tokens [25].

On the other hand, MARBERT uses the same network architecture as BERT but without the next sentence prediction (NSP) objective because it's trained only on tweet data, and tweets are short [26]. To fine-tune these models to deal with the classification task, we used the technique of freezing the entire models and then attached trainable, fully connected layers at the top head of the pre-trained models and use objective functions according to our needs. The number of neurons in that layer equals the classes we have in our dataset (two classes). Then a Softmax layer is applied over all these neurons, which is the final layer that computes the model probabilities. Figure 3 shows the required modification in BERT architecture to fine-tune the pretrained network for downstream tasks. We used AraBERT v2, an optimized and large version of AraBERT with a better vocabulary and more data [25].

HYPER-PARAMETER OPTIMIZATION

We conducted several experiments to optimize the hyperparameters to reach the best performance. Hyperparameters have two types: optimizers and model-specific hyperparameters [3]. Model-specific hyperparameters are the parameters related to the structure of the deep learning model, such as the number of layers and type of hidden nodes. Our model added a fully connected, dense layer with a Softmax classifier. On the other hand, the optimizer's hyperparameters contain the involved parameters in the learning process, such as epochs, batch size, and learning rate. The hyperparameters of these models are presented in Table 3.

Many optimizers' algorithms are available for training deep learning-based models, such as RMSProp, SGD, AdaGrad, and Adam [58]. In the proposed models, we used the Adam optimizer [59] with epsilon equal to 1e-8.

D. EVALUATION METRICS

The proposed system's evaluation will be done quantitatively to evaluate the performance and overall results. Evaluating ML and DL models is achieved by employing several metrics. The most widely used metrics are accuracy, precision, recall, and F1 score [25]. These measurements are outlined as follows: Accuracy is the total number of correctly predicted values and is computed by the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

where TP is the number of tweets identified as rumors correctly, FP is the number of tweets identified as rumors incorrectly. TN is the number of tweets that are identified as non-rumor correctly. FN is the number of tweets that are identified as non-rumor incorrectly. Precision represents the

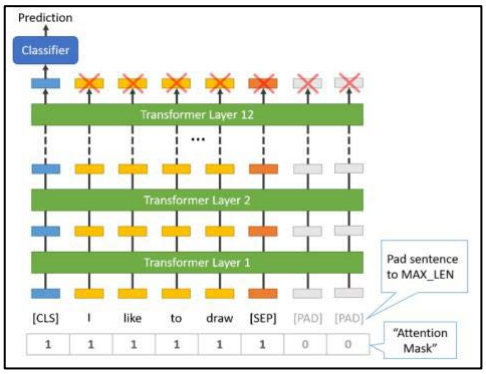


FIGURE 3. Top left node is responsible for classification that is used in the NLP task [30].

TABLE 3. Hyper-parameter setting.

Model	HYPER-PARAMETERS
AraBERT	Embedding size: 100 Batch Size: 40 Epochs: 8 Learning rate: 5e-5
MAEBERT	Embedding size: 100 Batch Size: 32 Epochs: 8 Learning rate: 5e-5

percentage of positively classified tweets that are correct. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall indicates the ability of the classifiers to classify all positive instances correctly. The formula computes it is the following:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The weighted harmonic mean of both precision and recall is defined as the F1-score and calculated by the formula:

$$F1 = \frac{2\,(Precision * Recall)}{Precision + Recall} \tag{4}$$

V. RESULTS

Table 4 shows the result of the proposed fine-tuned models AraBERT and MARBERT with the three datasets. The MARBERT model outperforms the AraBERT in the Fake News Detection and Arabic rumor-non-rumor tweets datasets with an accuracy of 0.78 and 0.97, respectively, while for AraBERT, the accuracy was 0.70 and 0.95. Whereas for the third dataset, COVID-19 misinformation, AraBERT outperformed MARABERT and achieved an accuracy of 0.90 compared to 0.88 for MARBERT.

Since two of the datasets (COVID-19 misinformation and Arabic rumor-non-rumor tweets) are imbalanced, to mitigate

TABLE 4. Results of the proposed models.

Model	Dataset	Acc.	Pre.	Rec.	F1
AraBERT	COVID-19 misinformation	0.90	0.81	0.75	0.78
	Fake News Detection	0.70	0.70	0.70	0.70
	Arabic rumor-non-rumor tweets	**0.95**	**0.94**	**0.93**	**0.93**
MAEBERT	COVID-19 misinformation	0.88	0.79	0.70	0.73
	Fake News Detection	0.77	0.77	0.77	0.77
	Arabic rumor-non-rumor tweets	**0.97**	**0.96**	**0.95**	**0.96**

the impact of this issue on the quality of the classification, two versions for each dataset were created: an undersampled and oversampled dataset. Table 5 displays the details of these datasets. The class ratio of rumor and non-rumor is approximately 1:1 in both undersampled and oversampled datasets, compared to 1:6 in the original one for the COVID-19 misinformation and 1:3 for the Arabic rumor-non-rumor tweets datasets. Table 6 shows the results after applying undersampling to the minority class. Results show that accuracy is decreased for all experiments, while the recall and F1 score is increased for training the COVID-19 misinformation dataset for both AraBERT and MARBERT models.

Table 7 illustrates the results after applying the oversampling to the majority class. The models show a significant improvement.

The Arabic rumor-non-rumor tweets dataset, which contains general tweets, was trained on both models, the AraBERT and MARBERT, and then validated using the datasets about COVID-19, which are the COVID-19 misinformation and Fake News Detection. Table 8 illustrates the best performance in terms of accuracy when validating the AraBERT and MARBERT using the COVID-19 dataset with scores of 0.52 and 0.53, respectively.

Next, we compared our model's result with the baseline models [50], [51], [55], and other Arabic rumor detection models mentioned in the literature; the comparison is illustrated in Table 9. These studies used both ML and DL techniques. The performance of the detection methods is compared in terms of accuracy and F1 score. The proposed model achieved the highest accuracy using MARBERT.

VI. DISCUSSION

In this study, we propose two rumor detection models using transformer-based models named AraBERT and MARBERT. Three Arabic public datasets were used, one is general about Arabic rumors, and the others are domain-specific about the COVID-19 pandemic. For each model, we fine-tuned the models and conducted experiments using the three datasets.

TABLE 5. Number of tweets after resampling.

Dataset	Original		Oversampled		Undersampled	
	Non-Rumor	Rumor	Non-Rumor	Rumor	Non-Rumor	Rumor
COVID-19 misinformation	7,475	1,311	7,475	7,866	1,311	1,311
Arabic rumor-non-rumor tweets	27,214	9,094	27,214	27,214	9,094	9,094

TABLE 6. Results after undersampling.

Model	Dataset	Acc.	Pre.	Rec.	F1
AraBERT	COVID-19 misinformation	0.80	0.80	0.80	0.80
	Arabic rumor-non-rumor tweets	**0.92**	**0.92**	**0.92**	**0.92**
MAEBERT	COVID-19 misinformation	0.75	0.77	0.75	0.75
	Arabic rumor-non-rumor tweets	**0.95**	**0.95**	**0.95**	**0.95**

TABLE 7. Results after oversampling.

Model	Dataset	Acc.	Pre.	Rec.	F1
AraBERT	COVID-19 misinformation	0.96	0.97	0.96	0.96
	Arabic rumor-non-rumor tweets	0.96	0.96	0.96	0.96
MAEBERT	COVID-19 misinformation	**0.98**	**0.98**	**0.98**	**0.98**
	Arabic rumor-non-rumor tweets	**0.99**	**0.99**	**0.99**	**0.99**

Two of these datasets are unbalanced. Therefore, resampling was applied to the datasets to mitigate the bias. In addition, we trained the models using the general dataset and validated them using the COVID-19-related datasets. All experiments' best results were obtained when the oversampling technique was applied, followed by training the original dataset. Additionally, the results showed that the models performed better when the undersampling method was not used. The results obtained from training the models on the Arabic rumor-non-rumor tweets dataset and validation on both COVID-19 misinformation and Fake News Detection datasets had the lowest results.

Table 9 illustrates the results of the proposed approaches with state-of-the-art methods for Arabic rumor detection. In [51], the study that provided the COVID-19 misinformation dataset, eight different traditional and deep machine learning models were used to detect rumors. The best result was achieved with a traditional classifier, namely, the Extreme Gradient Boosting (XGBoost), with an accuracy of 0.86. In this work, using the same dataset, AraBERT achieved 0.90 and MARBERT 0.88. The second dataset used in this work was obtained from [55], where traditional ML techniques were used to model the rumor detection task, and the best model achieved an F1 score of 0.87. Our proposed model outperformed this baseline model by 0.09.

Regarding the Arabic rumor-non-rumor dataset [48] used in this study, our results were 0.95 using AraBERT and 0.97 using MARBERT. The original study [48] applied both semi-supervised and unsupervised machine learning techniques to find rumors based on content and user features. The results indicated that the semi-supervised system, using a small base of labeled data, outperforms the supervised system and achieves 0.79 accuracy. In our work, we used only the tweets' content for the classification task. Thus, the unlabeled and duplicated data were excluded, which yielded a significant reduction in the size of the dataset, only 14% of the original dataset. The comparative analysis [53] used multiple Arabic transformers models; with AraBERT v2, the accuracy was 0.82, and 0.94 with MARBERT with gradual unfreezing, special learning rate, and learning rate scheduling. While in this work, a constant learning rate was used. These results reinforce the studies that confirmed that transformers-based

TABLE 8. Results of validating models on covid-19 datasets.

Model	Dataset	Acc.
AraBERT	COVID-19 misinformation	0.48
	Arabic rumor-non-rumor tweets	**0.52**
MAEBERT	COVID-19 misinformation	0.47
	Arabic rumor-non-rumor tweets	**0.53**

TABLE 9. Comparison of the proposed approach with the state-of-the-art methods.

Reference	Method Used	Accuracy	F1
[47]	Machine Learning, TF-IDF	0.67	–
[49]	Machine Learning	0.86	0.79
[51]	Deep Learning, eight different traditional and deep machine learning models	0.86	0.54
[52]	DL	0.81	0.81
[55]	Traditional ML	–	0.87
[48]	ML, semi-supervised expectation–maximization (E–M), and supervised GNB	0.79	0.80
[53]	Traditional ML, and DL	0.97	0.95
Proposed	Transformer-based using MARBERT	**0.97**	**0.96**

models outperform and enhance the results of other deep learning-based models [24], [53]. Transformer-based models have several factors that can explain these results; one of them is the extensive language knowledge gained by the transformers by training them on language modeling objectives [53].

Transformers-based models for English rumor detection mainly achieve better results than the Arabic models [19], [39], [58]. Models that rely on Arabic content face many challenges. Arabic has rich and complex grammar with a massive vocabulary. Also, it has various dialects people use in their life and on social media posts. This variety produces many new words in the language [53]. Another challenge that increases the complexity is that many Arabic vocabularies have different meanings based on diacritics and context. Furthermore, it contains many grammatical rules that change the words' purpose and shape [45]. Regardless of all these challenges, the proposed model achieved reasonably high performance, as shown in Table 4.

In the resampling experiments, the models achieved better with oversampling the rumor class, which is the lowest class on both datasets. The undersampling of the majority class (non-rumor) affected the results negatively. To evaluate the generalizability of the models, we trained the models on Arabic rumor-non-rumor tweets. We then evaluated them on COVID-19 misinformation and Fake News Detection datasets, as shown in Table 8. The models perform differently depending on the dataset. The results show that the accuracy is between 0.25 and 0.45, lower than in the other experiments. This indicates that the tweets' domain greatly influences the classifier's performance. The training data contains tweets in various fields and topics but no content related to the pandemic or COVID-19. These results could be improved by training the models on datasets in a similar domain to the test dataset.

VII. CONCLUSION AND FUTURE WORK

Rumors took center stage during the COVID-19 pandemic, where fake news, conspiracy theories, and false medical advice circulated on social media like never before. This study aimed to create an accurate model using advanced transformer models to filter genuine Arabic posts from rumors. Many contributions to this domain were found in the literature adopting different approaches, including traditional ML, DL, and the pre-trained transformer models. Many Arabic text classification studies investigated standard ML and DL models in the rumor detection task. Few studies applied transformers to Arabic NLP applications. In this study, two Arabic transformer-based deep learning models, AraBERT and MARBERT, were used to develop an Arabic rumor detection model. The results reached outperformed previous deep learning models applied to the same dataset. Our results support the findings of other studies that used transformers, which stated that transformer-based models yield better results than other deep learning-based models [24], [53].

We considered publicly available datasets not only to reduce the time and effort needed to gather and label our dataset but to enable the comparison with a baseline model. However, carefully examining the posts and their labels in these datasets revealed some inaccurate labels. Inconsistent labeling may explain inefficiencies in machine learning models. Inline future directions with this work include constructing more extensive and accurate datasets, extending the model to detect rumors in different domains, and experimenting with other transformer-based models with multiple datasets. We hope that this effort will complement the pursuit of developing a generalizable model that would fight fake news on social media to protect people from falling into misleading information.

REFERENCES

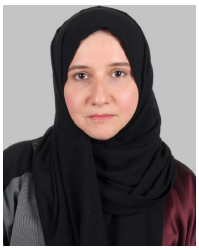
[1] J. G. W. Allport and L. Postman, *The Psychology of Rumor*. Oxford, U.K.: Henry Holt, 1947.
[2] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102018.
[3] A. R. Pathak, A. Mahajan, K. Singh, A. Patil, and A. Nair, "Analysis of techniques for rumor detection in social media," *Proc. Comput. Sci.*, vol. 167, pp. 2286–2296, Jan. 2020.
[4] T. Granskogen and J. A. Gulla, "Fake news detection: Network data from social media used to predict fakes," in *Proc. CEUR Workshop*, vol. 2041, no. 1, 2017, pp. 59–66.
[5] P. Domm. (2013). False Rumor of Explosion at White House Causes Stocks to Briefly Plunge; AP Confirms Its Twitter Feed Was Hacked. CNBC, Mearket Insider. Accessed: Jan. 2022. [Online]. Available: https://www.cnbc.com/id/100646197
[6] *Who We Are*, N. R. Commission, 2021.
[7] J. D. Kelleher, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2019.
[8] D. A. Bashar, "Survey on evolving deep learning neural network architectures," *J. Artif. Intell. Capsul. Networks*, vol. 2019, no. 2, pp. 73–82, 2019.
[9] L. Tian, X. Zhang, Y. Wang, and H. Liu, "Early detection of rumours on Twitter via stance transfer learning," *Advances in Information Retrieval* (Lecture Notes in Computer Science), vol. 12035. Cham, Switzerland: Springer, 2020.
[10] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K. F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 3818–3824.
[11] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 1751–1754.
[12] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proc. Conf. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL) (Long Papers)*, vol. 1, 2017, pp. 708–717.
[13] X. D. Zhang, *A Matrix Algebra Approach to Artificial Intelligence*. Cham, Switzerland: Springer, 2020.
[14] Y. Bengio, "Machines who learn," *Sci. Amer.*, vol. 314, no. 6, pp. 44–51, 2016.
[15] K. R. Chowdhary, *Fundamentals of Artificial Intelligence*. Cham, Switzerland: Springer, 2020.
[16] R. Alharthi, A. Alhothali, and K. Moria, "A real-time deep-learning approach for filtering Arabic low-quality content and accounts on Twitter," *Inf. Syst.*, vol. 99, Jul. 2021, Art. no. 101740.
[17] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4593–4601.
[18] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4487–4496.
[19] S. M. S.-U.-R. Shifath, M. F. Khan, and M. S. Islam, "A transformer based approach for fighting COVID-19 fake news," *arXiv*, pp. 1–9, 2021.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[21] M. Wiercioch and J. Kirchmair, "Dealing with a data-limited regime: Combining transfer learning and transformer attention mechanism to increase aqueous solubility prediction performance," *Artif. Intell. Life Sci.*, vol. 1, Dec. 2021, Art. no. 100021.
[22] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (NAACL HLT)*, vol. 1, 2019, pp. 4171–4186.
[23] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 4944–4953.
[24] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 328–339.
[25] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools, Shared Task Offensive Lang. Detection*, 2020, pp. 9–15.
[26] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2021, pp. 7088–7105.
[27] T. Rus, "Natural language processing," in *Encyclopedia of Sciences and Religions*. Cham, Switzerland: Springer, 2013.
[28] D. Namly, K. Bouzoubaa, A. El Jihad, and S. L. Aouragh, "Improving Arabic lemmatization through a lemmas database and a machine-learning technique," in *Recent Advances in NLP: The Case of Arabic Language*. Cham, Switzerland: Springer, 2020, pp. 81–100.
[29] M. Alkhatib, M. El Barachi, and K. Shaalan, "An Arabic social media based framework for incidents and events monitoring in smart cities," *J. Cleaner Prod.*, vol. 220, pp. 771–785, May 2019.
[30] F. R. Alharbi and M. B. Khan, "Identifying comparative opinions in Arabic text in social media using machine learning techniques," *Social Netw. Appl. Sci.*, vol. 1, no. 3, pp. 1–13, Mar. 2019.
[31] G. Amoudi, R. Albalawi, F. Baothman, A. Jamal, H. Alghamdi, and A. Alhothali, "Arabic rumor detection: A comparative study," *Alexandria Eng. J.*, vol. 61, no. 12, pp. 12511–12523, Dec. 2022.
[32] Y. Yuan, Y. Wang, and K. Liu, "Perceiving more truth: A dilated-block-based convolutional network for rumor identification," *Inf. Sci.*, vol. 569, pp. 746–765, Aug. 2021.
[33] R. Sicilia, S. Lo Giudice, Y. Pei, M. Pechenizkiy, and P. Soda, "Twitter rumour detection in the health domain," *Expert Syst. Appl.*, vol. 110, pp. 33–40, 2018.
[34] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular Twitter threads," Tech. Rep., 2017.
[35] Q. Huang, C. Zhou, J. Wu, L. Liu, and B. Wang, "Deep spatial–temporal structure learning for rumor detection on Twitter," *Neural Comput. Appl.*, vol. 17, pp. 1–8, Aug. 2020.
[36] K. Tu, C. Chen, C. Hou, J. Yuan, J. Li, and X. Yuan, "Rumor2vec: A rumor detection framework with joint text and propagation structure representation learning," *Inf. Sci.*, vol. 560, pp. 137–151, Jun. 2021.
[37] A. P. B. Veyseh, M. T. Thai, T. H. Nguyen, and D. Dou, "Rumor detection in social networks via deep contextual modeling," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2019, pp. 113–120.
[38] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, "Evaluating deep learning approaches for COVID-19 fake news detection," in *Proc. Int. Workshop Combating Online Hostile Posts Regional Lang. During Emergency Situation*, 2021, pp. 153–163.
[39] B. Chen, "Transformer-based language model fine-tuning methods for COVID-19 fake news detection," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation* (Communications in Computer and Information Science), vol. 1402. Cham, Switzerland: Springer, 2021.
[40] M. Alruily, O. R. Shahin, H. Al-Mahdi, and A. I. Taloba, "Asymmetric DNA encryption and decryption technique for Arabic plaintext," *J. Ambient Intell. Hum. Comput.*, vol. 2, pp. 1–17, Apr. 2021.
[41] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, nos. 412–420, 1997, p. 35.
[42] S. M. Alzanin, A. M. Azmi, and H. A. Aboalsamh, "Short text classification for Arabic social media tweets," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 9, pp. 6595–6604, Oct. 2022.
[43] A.-S. Hissah and H. Al-Dossari, "Detecting and classifying crimes from Arabic Twitter posts using text mining techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 10, pp. 1–11, 2018.
[44] A. M. Bdeir and F. Ibrahim, "A framework for Arabic tweets multi-label classification using word embedding and neural networks algorithms," in *Proc. 2020 2nd Int. Conf. Big Data Eng.*, 2020, pp. 105–112.
[45] A. Alotaibi, "Racism detection in Twitter using deep learning and text mining techniques for the Arabic language," Tech. Rep., 2020, pp. 161–164.
[46] H. Chouikhi, H. Chniter, and F. Jarray, "Arabic sentiment analysis using BERT model," *Commun. Comput. Inf. Sci.*, vol. 1463, pp. 621–632, Sep. 2021.
[47] A. Y. M. Floos, "Arabic rumors identification by measuring the credibility of Arabic Tweet content," in *Media Controversy: Breakthroughs in Research and Practice*. Hershey, PA, USA: IGI Global, 2020, pp. 236–248.
[48] S. M. Alzanin and A. M. Azmi, "Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowl.-Based Syst.*, vol. 185, Dec. 2019, Art. no. 104945.
[49] H. Himdi, G. Weir, F. Assiri, and H. Al-Barhamtoshy, "Arabic fake news detection based on textual analysis," *Arabian J. Sci. Eng.*, pp. 1–17, 2022.
[50] A. Khalil, M. Jarrah, M. Aldwairi, and Y. Jararweh, "Detecting Arabic fake news using machine learning," in *Proc. 2nd Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA)*, Nov. 2021, pp. 171–177.
[51] 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," Tech. Rep., 2021.
[52] S. E. Sorour and H. E. Abdelkader, "AFND: Arabic fake news detection with an ensemble deep CNN-LSTM model," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 14, pp. 5072–5086, 2022.
[53] M. Al-Yahya, H. Al-Khalifa, H. Al-Baity, D. AlSaeed, and A. Essam, "Arabic fake news detection: Comparative study of neural networks and transformer-based approaches," *Complexity*, vol. 2021, pp. 1–10, Apr. 2021.
[54] E. M. B. Nagoudi, A. Elmadany, M. Abdul-Mageed, T. Alhindi, and H. Cavusoglu, "Machine generation and detection of Arabic manipulated and fake news," Tech. Rep., 2020, pp. 1–15.
[55] A. R. Mahlous and A. Al-Laith, "Fake news detection in Arabic tweets during the COVID-19 pandemic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 1–10, 2021.
[56] O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, and N. Habash, "CAMeL tools: An open-source Python toolkit for Arabic natural language processing," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 7022–7032.
[57] K. Darwish and H. Mubarak, "Farasa: A new fast and accurate Arabic word segmenter," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 1070–1074.
[58] Z. Wu, D. Pi, J. Chen, M. Xie, and J. Cao, "Rumor detection based on propagation graph neural network with attention mechanism," *Expert Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113595.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

NAELAH O. BAHURMUZ received the bachelor's degree in information technology, in 2015. She is currently pursuing the master's degree in information systems with the Faculty of Computing and Information Technology, King Abdulaziz University. Her research interests include machine learning and deep learning.

GHADA A. AMOUDI received the Ph.D. degree in computer science from Dalhousie University, Canada, in 2016. She is currently an Assistant Professor with the Faculty of Computing and Information Technology, King Abdulaziz University. She has been involved in academic activities, since 2005, and worked as an Educator, a Lecturer, and a Researcher. She has coauthored an article in the imaging deep learning domain titled *Deep Learning Approaches for Detecting COVID-19 From Chest X-Ray Images: A Survey* and another recent one in the deep learning NLP domain titled *Arabic Rumor Detection: A Comparative Study*. Her research interests include applying AI to Arabic text mining, health domains, and social media. She is the Co-Founder and a member of the Director Board of the Saudi Artificial Intelligent Association.

FATMAH A. BAOTHMAN received the Ph.D. degree in modern artificial intelligence from the School of Computing and Engineering, the University of Huddersfield, in 2003. She is currently a Faculty Member with KAU. She is the Founder and the Board President of AI Society in Saudi Arabia, the first woman internationally awarded an AI prize from USA and U.K. She is the first AI Woman in the middle east and the President of the Saudi Engineering Society, IEEE Chapter at KAU. She gained several trophies from different ministries and governmental bodies for her input on various occasions and programs. She held several administrative leading positions inside and outside KAU. Among these positions are the apple center manager, a GM IT security, the director for KAEC–Educational Sector, deputy director of the IT Center, KAU, the Chairperson of the IEEE Women in Engineering Western Region, and the President of Women Engineers Committee at Saudi Council of Engineering. She participated in forming technology strategies with the University and the National Strategic IT Plan for Information technology with KACST. She is authoring several books and supervised many applied technology projects. She taught and introduced workshops in artificial intelligence, emergent technology, the IoT, data mining, blockchain, AI and entrepreneurship, and IR4. She translated an AI book into Arabic and coined several technical terms. She is the first Saudi to join the Artificial Intelligence Global Educational Theater. She led the project successfully in two academic institutes. She also participated in the 20th AI Laboratory Anniversary and the 50th AI Anniversary among AI Global Scientists.



AMANI T. JAMAL received the master's and Ph.D. degrees from Concordia University, Montreal, Canada. She is currently an Associate Professor with the Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University. Her current research interests include natural language processing and computer vision related to Arabic text and historical documents. She is the Co-Founder and a member of the Director Board of the Saudi Artificial Intelligent Association.

HANAN S. ALGHAMDI received the Ph.D. degree in computer science from the University of Surrey, Guildford, U.K., in 2018. She is currently an Assistant Professor at King Abdulaziz University, Jeddah, Saudi Arabia. She has published several papers in the field of medical images analysis; examples include *Measurement of Optical Cup to Disk Ratio in Fundus Images for Glaucoma screening*, *Ensemble Learning Optimization for Diabetic Retinopathy Image Analysis*, *Automatic Optic Disc Abnormality Detection in Fundus Images: A Deep Learning Approach*, and *Deep learning approaches for detecting COVID-19 From Chest X-Ray Images: A Survey*. Her current research interests include medical image analysis, explainable artificial intelligence, deep learning reinforcement learning, generative adversarial networks, evolutionary algorithms, and optimization. She is involved in several research activities and has participated as a reviewer for multiple journals and conferences locally and internationally. Examples include IEEE Access journal, *Scientific Reports* journal, International Conference on Computational Intelligence and Intelligent Systems, CIIS, International Conference on Machine Learning and Human–Computer Interaction, MLHMI, Asia Conference on Algorithms, Computing and Machine Learning, *CACML*, and the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI. Regarding professional experience, she worked as the Strategic Communication Manager at the Savola Group, Saudi Arabia, and a Demonstrator at the University of Surrey, U.K.

AREEJ M. ALHOTHALI received the master's and Ph.D. degrees in computer science (artificial intelligence) from the University of Waterloo, Canada, in 2017. She is currently an Assistant Professor with the Faculty of Computer Science and Information Technology, King Abdulaziz University. Her research interests include machine learning, deep learning, natural language processing, computer vision, intelligent agent systems, and affective computing.