Fake News Detection Using Enhanced BERT

Shadi A. Aljawarneh and Safa Ahmad Swedat

***Abstract*— Since there are so many users on social media, who are not qualified to report news, fake news has become a major problem in recent years. Therefore, it is crucial to identify and restrict the dissemination of false information. Numerous deep learning models that make use of natural language processing have yielded excellent results in the detection of fake news. bidirectional encoder representations from transformers (BERT), based on transfer learning, is one of the most advanced models. In this work, the researchers have compared the earlier studies that employed baseline models versus the research articles where the researchers used a pretrained model BERT for the detection of fake news. The literature analysis revealed that utilizing pretrained algorithms is more effective at identifying fake news because it takes less time to train them and yields better results. Based on the results noted in this article, the researchers have advised the utilization of pretrained models that have already been taught to take advantage of transfer learning, which shortens training time and enables the use of large datasets, as well as a reputable model that performs well in terms of precision, recall, as well as the minimum number of false positive and false negative outputs. As a result, the researchers created an improved BERT model, while considering fine-tuning it to meet the demands of the fake news identification assignment. To obtain the most accurate representation of the input text, the final layer of this model is also unfrozen and trained on news texts. The dataset used in the study included 23 502 articles of fake news and 21 417 items of actual news. This dataset was downloaded from the Kaggle website. The results of this study demonstrated that the proposed model showed a better performance compared with other models, and achieved 99.96% and 99.96% in terms of accuracy and *F*1 score, respectively.**

***Index Terms*— Bidirectional encoder representations from transformers (BERT), fake news, pretrained model, social networks.**

I. Introduction

THE development of social media networks in recent years has made it easier for people to communicate with one another. Social media has become the more popular platform for communication and information transfer due to its simplicity of use, quick expansion rate, and low cost [1]. Social media users share data and keep up with the latest trends by using sites like Facebook and Twitter. However, a lot of this material is questionable and unreliable, which is called false news [10].

The spread of false information causes serious issues that could have an impact on people's fates. Fake news has also impacted the 2016 U.S. presidential election campaign [1], [12]. In the 5 months before the election, 30 million tweets were accumulated. A 25% of these tweets contained biased or fake news. Most of this fake news was circulated by Trump supporters, who also used it to influence the activities of Clinton supporters and assist Trump in winning the election [13]. The researchers' perspectives on the significance of determining if the information is accurate are also changing. As the content creator wants to deceive the reader, they will write the news in a way that makes it difficult for the reader to tell the difference between fake and real news, which makes the detection of fake news challenging [11]. Furthermore, the information gleaned through social media is vast, frequently anonymous, user-generated, and noisy. These difficulties have led to the need to use more sophisticated techniques for detecting false news [11].

A few studies [5], [6], [9] focused on the false news identification field and displayed good accuracy. Despite this, most of these studies employ both deep learning and traditional machine learning methods. In this study, the researchers have compared the pretrained bidirectional encoder representations from transformers (BERT) model with the baseline models for detecting false news. BERT is a deep learning method based on transformers and uses transfer learning. It has been developed by Google for natural language processing (NLP) [2]. A sizable corpus of unlabeled text has been used to pretrain BERT. The training texts consisted of Book Corpus (800 million words) and the entirety of Wikipedia, totaling 2 500 000 000 words. The model was trained for performing two tasks, i.e., next sentence prediction (NSP) and masked language modeling (MLM) [2]. BERT uses the Transformers to learn how words in a text relate to one another in context. Transformers, on the other hand, consisted of two distinct mechanisms: a decoder to make predictions for the task as output and an encoder to receive text input [14]. Given that the main objective of BERT is to produce a language model, it only comprises the encoder mechanism. BERT consists of 12 encoders and 12 self-attention heads. The input for each encoder is 768 tokens, which are initially embedded into the vectors and then processed by a neural network. The result is a series of vectors having the same size as the input, where every vector corresponds to a token from the input with the same index [15]. To carry out experiments and identify fake news using the dataset retrieved from the Kaggle website that is nearly balanced, this study makes use of the advantages of the enhanced BERT model. These experiments showed a better performance than the models used in the earlier studies and showed an accuracy and $F1$ score of 99.96% and 99.96%, respectively. Our research question is how to improve

Manuscript received 16 July 2022; revised 9 September 2022 and 11 October 2022; accepted 18 November 2022. *(Corresponding author: Shadi A. Aljawarneh.)*

The authors are with the Faculty of Computer and Information Technology, Jordan University of Science and Technology, Irbid 22110, Jordan (e-mail: saaljawarneh@just.edu.jo; saswedat20@cit.just.edu.jo).

Digital Object Identifier 10.1109/TCSS.2022.3223786

the BERT model to achieve the highest accuracy, $F1$ score, precision, recall, and least error rate in terms of fake news detection.

The main contributions of this study are listed shortly in these points.

1) Improving a big dataset that has long text for each record.
2) Utilizing the transfer learning and pretrained models to reduce the training time and make use of a large dataset.
3) Enhancing a BERT model which takes into consideration the sequential nature of the input data and the contextual meaning of the texts.
4) Achieving a high recall, precision, and the least amount of false positive and false negative results.

The remaining study is organized in the following manner. Section II presents a discussion of a few related studies. In Section III, the researchers have discussed the gaps existing in the studies and the implications of this study. In Section IV, the researchers have discussed the methodology used in this study. Section V has presented the results of all experiments conducted in this study. Finally, Section VI has mentioned the conclusions of the study.

II. Related Work

The impact of fake news on people and society makes it one of the most essential research topics [5], [10]. To distinguish fake news from legitimate news, the researchers in [3] extracted information from texts using machine learning models and techniques. They evaluated their dataset using K-nearest neighbor, Naive Bayes, support vector machine (SVM), random forest, and XGBoost models. Though their models performed well, they were not effective with large datasets. In addition, the context of the statements was not considered. The context meaning of sentences or even the individual words is a critical determinant in any NLP process as it impacts the overall meaning of the texts and could confuse the classification model [11], [12]. For instance, the term “book” denotes something to read in the statement “I have read a book this weekend.” Whereas it also means to “reserve a flight” in the statement, “I have to book a flight before next Friday.”

A more complex model called FNDNet was proposed by Kaliyar et al. [4]. FNDNet is a system that uses deep convolutional neural networks (CNNs) to identify false information that circulated during the 2016 U.S. Presidential Election. They applied the GLoVe word embedding technique and compared their model to numerous baseline models, including random forest (RF), decision tree, multinomial Naive Bayes, and K-nearest neighbor. When compared with the baseline models, their model performed exceptionally well. Despite this, their strategy does not rely on context, content, or temporal-level data.

To learn the order and correlation between the news content and to explain the prediction, Shu et al. [5] suggested a model based on the sentence-comment and co-attention. Their model, which was created using BiLSTM, was applied to a sizeable dataset acquired from PolitiFact and GossipCop. This model showed better performance, with regard to the accuracy rate and $F1$ score (0.904 and 0.928, respectively), using the PolitiFact dataset. The model did not perform as well on the second dataset as it did on the first, indicating that it did not behave consistently across different datasets. Additional side information was needed to find the comprehensible remarks that will strengthen the model.

To categorize false Bengali text in a dataset, Mugdha et al. [6] examined various supervised models. The best performance was displayed by the Gaussian Naive Bayes (GNB) algorithm, which uses an extra tree classifier to select the best features and text features based on term frequency–inverse document frequency (TF-IDF). This model fared better than the other models and displayed an accuracy rate of 87.42%. However, their study was restricted to using only the news headlines, not the entire article. Another study [7] presented a complete analysis to assess the effectiveness of 23 supervised machine learning models. The results indicated that the decision tree generated the best results. They established that deep learning models outperform the conventional machine learning models.

Singh [8] employed three models—the CNN, recurrent neural network (RNN), and artificial neural network—to illustrate the advantages of deep learning. He utilized four forms to represent the vector space, i.e., Doc2Vec, Word2Vec, TF–IDF, and one-hot vector encoding. The models were analyzed using two datasets, i.e., LIAR and Kaggle, and the TF–IDF vector representation showed a better performance using the Kaggle dataset. The dataset only contains short sentences; hence it is impossible to generalize the findings to all sentence lengths.

The two classification models and user-level security protocol comprise three components of HRSP. The classification of contents is one of the models. This methodology aims to categorize the content into three categories: hate speech, benign speech, and inappropriate speech. The seven machine learning algorithms—KNN, J48, SVM, NB, R-F, logistic regression (LR), and D-Table—were used in their study. They used a dataset of 22 000 tweets to test their algorithms, and their overall accuracy was 84.4%. The sole drawback of this study is that it only analyses the textual data, although online social networks contain a plethora of data.

Most studies on the topic of false news identification rely on conventional machine learning and deep learning methods. However, they are not strong enough. To save training time and develop more reliable and trustworthy models, researchers began to focus on pretrained models. To identify false information during the COVID-19 epidemic, some researchers [2] presented a transformer-based model for processing the explainable natural language using the BERT model and its modifications. They used SHAP (Shapley Additive exPlanations) to enhance the model’s functionality. To explain the results of the BERT model, SHAP assigns a significant value to each characteristic that is associated with a certain prediction. They tested their model using two sizable datasets and noted outstanding accuracy results of 0.972 and 0.938. However, most COVID-19 statements that might be confirmed are not included in their study, which suggests that these datasets may be deceptive and not accurately reflect the population. In addition, the study is restricted to identifying false information regarding COVID-19, and other pieces of knowledge are not used to predict the behavior of the model.

The dissemination of false information is not limited to the English language; Arabic is also affected. Due to a lack of resources and databases, detecting fake news in Arabic is a difficult undertaking [18], [19]. Harrag and Djahli [20] used the Arabic Fact-Checking and Stance Detection Corpus and said it is the only dataset that they could find in the Arabic language and supports the fake news detection problem. Another issue in detecting fake news in the Arabic language is the dialect variety. Hanen Himdi et al. [21] had to collect their dataset in three different dialects to deal with this problem.

The AraCOVID19-MFH dataset was published and manually annotated by Ameur and Aliane [22] for detecting fake news. They employed three baseline pretrained models in addition to two transformer models that were further pretrained on the COVID-19 dataset to familiarize them with COVID-19 terms, and then they applied them to the goal of identifying fake news. The model that was trained on COVID-19 words produced the greatest results, with a weighted F-score of 95.78%. Even though the dataset is small and could only be used for identifying the false news in COVID-19 news, it showed high accuracy.

For identifying bogus news in Arabic, Yahya et al. [23] compared the performance of the transformer-based (ArBert, ArElectra, AraBERT v1, AraBERT v2, AraBERT v02, QARiB, and MarBert) and neural network-based [CNN, RNN, and gated recurrent unit (GRU)] models. They utilized four datasets for evaluating the different models. They demonstrated that the transformers-based models showed a better performance than the conventional neural networks. The transformer-based QARiB model showed the highest $F1$ score of 95% and displayed the best performance. Despite the impressive outcome, their dataset was limited and contained repeated tweets. It also has noise and unclassifiable texts.

Mahlous and Al-Laith [24] proposed a novel dataset for detecting fake news. After cleaning, the dataset included 1537 manually annotated tweets and 34 529 automatically annotated tweets. Only machine learning algorithms, including Naive Bayes, LR, RF bagging model, multilayer perceptron (MLP), SVM, and eXtreme Gradient Boosting Model (XGB), were used to evaluate the datasets. The LR classifier showed a better performance than other classifiers for both datasets, with regard to the $F1$ score of 87.8% for the manually-annotated dataset, using the TF–IDF feature; and 93.3% for the automatically annotated dataset with the count vector as the feature. Although the results were excellent, the dataset used was very small and included different dialects, which made it difficult to determine the good features for classification. In addition, it was affected due to language mistakes.

For detecting fake news, Nassif et al. [25] assembled their dataset. They also utilized a different dataset that was released by Kaggle and was translated from English to Arabic using the Google Translate feature. To examine both datasets, they used eight transformer-based models designed for Arabic. Their findings demonstrated that working with the translated text was not similar in comparison to working on data that was originally written by the native speakers. The ARBERT model yielded the best results, with an E score of 98.9%, using their dataset. Their only drawback was a small dataset.

III. Conceptual Comparison

After comparing the different models, the following perspectives were derived in this study: 1) because fake news has such a negative influence on people and society, it must be prevented from spreading readily through social media networks; 2) researchers must use more reliable techniques to intercept any erroneous or misleading information that might be put on social media before users see it; 3) to train a model from scratch and produce a decent model with an accurate output, a large dataset and a lot of time are required for understanding the good correlation between the input characteristics; and 4) while several studies on false news identification have performed well in terms of accuracy and $F1$ scores, their studies were limited to short texts and small datasets because they used conventional machine learning and deep learning models. Based on these perspectives and the conclusions derived from other studies, the researchers in this study strived to take advantage of transfer learning and use pretrained models, which reduces training time and enables the use of large datasets, for generating a reliable model that displays a high precision, recall, and the least amount of false positive and false negative results.

Table I describes, in brief, all related studies based on the dataset that was used, either big or small, representing the population or not and whether it made the model more weak or robust. It is clear from Table I that most studies were made on small datasets where the model's behavior is not predicted or not efficient with big datasets. Moreover, some studies are limited to a specific type of datasets. In addition, some datasets have noise or are not clean.

Table II shows a brief description of the models used in every study, the accuracy rate that was achieved and the limitations of every study. It seems that most models have high accuracy or F-score rate but it lacks reliance on the content, context, or temporal-level information.

After comparing the various concepts, a new BERT model has been proposed in this study for fulfilling the study objectives.

IV. Methodology

A. Dataset and Preprocessing Steps

The study's dataset was downloaded from the Kaggle website [16]. The downloaded folder included two files, one of which contained Fake news, and the second folder contained True news. Both files included four columns: the Title column with article titles, the Text column with the articles themselves, the Subject column with the article's subject, and the Date column with the publication date for every article.

The dataset must be preprocessed before being used in the experiments. Each file contains data without assigned labels, therefore to identify false news, a column of individual values is added to the fake news file. In addition, a column of zeros is

TABLE I

LIMITATIONS OF THE STUDIES BASED ON THEIR DATASETS

Ref. No.	Small Dataset	Big Dataset	Limitations
[2]		√	Dataset may not representative
[3]	√		Model is not efficient with big data
[4]	√		Model behaviour is not predicted with big data
[5]		√	Model not robust with different type of datasets
[6]	√		Dataset is very small and not representative
[7]	√	√	Not same behaviour for different sizes of datasets
[8]	√		Dataset consist of short sentences
[9]	√		Behaviour of model is not predicted for bigger dataset
[22]	√		Dataset is limited for fake news detection in Covid19 news
[23]	√		Dataset has repetition of tweets and suffers from noise
[24]	√		Dataset consist of various dialects and it suffers of language mistakes
[25]	√		The dataset is small

TABLE II

LIMITATIONS OF THE STUDIES BASED ON THE MODELS THAT WERE USED AND THEIR ACCURACY RATES

Ref. No.	Used Model	Accuracy	Limitation
[2]	Transformer-based (BERT)	**97.20%**	Model is limited for COVID-19 dataset
[3]	Baseline machine learning model	**86.00%**	Model did not consider the contextual meaning
[4]	FNDNet with GLoVe pre-trained word embedding	**98.36%**	Model lacks reliance on the content, context, and temporal level information.
[5]	Bi-LSTM	**90.40%**	Model needs more side information to discover the explainable comments
[6]	GNB	**87.42%**	Model built based on news headlines only
[7]	Decision Tree	**96.80% for big data 68.00% for small data**	Model behaviour is different for different datasets
[8]	TFIDF with CNN	**96.89%**	Poor behaviour with bigger dataset
[9]	Random Forest	**84.40%**	-
[22]	Transformer-based models	**95.78%** in term of Weighted F-score	Model is limited for COVID-19 dataset
[23]	Recurrent neural networks and Transformer-based models	**95.00%** in term of F1 score	-
[24]	Machine Learning techniques	**93.30%** in term of F1 score	-
[25]	Transformer-based models	**98.90%** in term of F1 score	-

inserted in the true news file to confirm if the news is authentic (true).

After labeling the data in the two files, they were concatenated into one file and its contents were randomly ordered to prevent sequencing data with the same label, in the training phase. The file is then cleared of null values. For training purposes, only the column in the articles is used because, in addition to the label column, it contains the primary factor that can be used to decide whether the news is true or false. As a result of their length, the subject column, date, and title both contain less information related to the news than the actual article itself, and they have all been removed.

By the time the data had been cleaned, the dataset had 44 898 records; 23 481 of those were fake, while 21 417 were true, indicating that the data were nearly balanced and that no sampling approaches were required.

B. Functionality Overview of the Proposed Model

In this section, the proposed model steps are discussed briefly.

First: The dataset is first processed as described in the Methodology section.

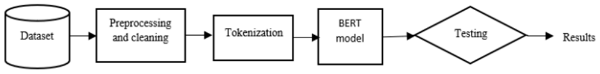

Fig. 1. Steps used in this study.

Second: The BERT Tokenizer is used to extract the token sequence for each text in the dataset.

Third: Each token is transformed into a vector representation that BERT can accept.

Fourth: For training purposes, BERT is then fed the input representation for every sequence.

Finally: The produced model is tested using the test data and getting the final result. Fig. 1 shows the process used in this article.

C. Enhanced BERT Model

Since text data are sequential, the training model must accommodate this nature, where the features' order is a key consideration [17]. In addition, as was already established, every text categorization task must consider the text's context [11], [12]. Thus, the pretrained transformers model is advantageous for this study (BERT). BERT is an encoder-based, multilayered, and bidirectional transformer. BERT has two different variants, where the small model is called the base, while the other one is large. BERT-base is made up of 12 layers of transformers (encoder). Every encoder is made up of a stack of six similar layers, and every layer is made up of two sublayers. The first layer of each encoder has a multihead self-attention mechanism (12 attention heads), while the second layer is a simple position-wise feed-forward network having a residual connection around both the sublayers, followed by a normalization layer. A series of tokens with a size of 768 is fed into the transformer encoder as input. These tokens are first embedded into the vectors and then processed by the neural network to generate the output. The output includes a series of 512-sized vectors, each of which represents an input token having the same index [14], [15].

BERT is built to handle a variety of downstream jobs by allowing the input representation to be represented in a single token sequence for both one and two sentences having no ambiguity. The sentence may not even be a linguistic one; it may be related to a block of text. The BERT model's "sequence" refers to the input token, which could be one or two sentences combined. BERT employs 30 000 tokens of language with WordPiece embeddings [26]. The first token in each sequence must be a special classification token (CLS). For classification tasks, the aggregate sequence representation corresponding to (CLS) is derived from the final hidden state. Sentences are grouped together into a single sequence and distinguished in one of two ways: The sentences are first divided using a special token (SEP). Second, a learned embedding is added to every token in the sequence for indicating if they belong to sentence A or sentence B. For every token, another embedding known as a positional embedding is added to identify its location in the sequence. The final hidden vector of [CLS] is denoted by C, the input embedding is designated as E, and the final hidden vector for the ith input token is denoted by T_i as shown in Fig. 2 [15].

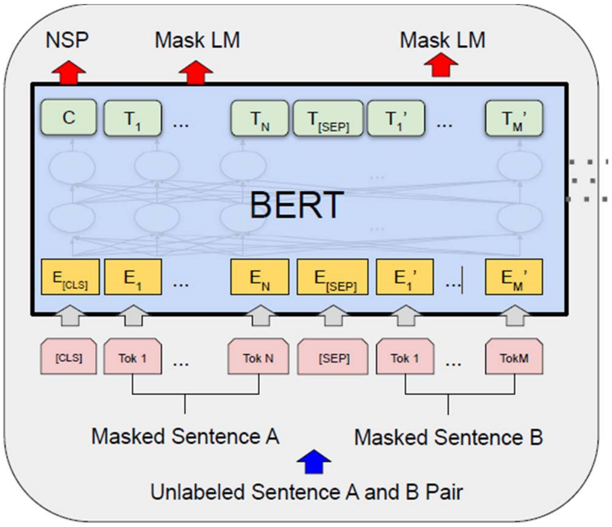

Fig. 2. General pretraining procedures used in BERT.

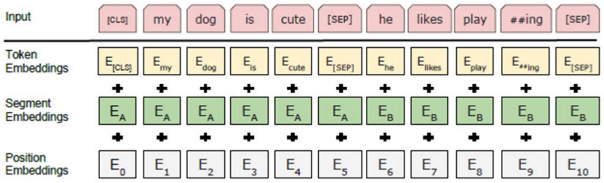
Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
E[CLS] Emy Edog Eis Ecute E[SEP] Ehe Elikes Eplay E##ing E[SEP]
Segment Embeddings
EA EA EA EA EA EA EB EB EB EB EB
Position Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10

Fig. 3. BERT input representation.

By adding the corresponding token, segment, and position embeddings, the input representation for a particular token is determined. This design is described in Fig. 3 [15].

Traditional language models only read the input text in a single direction, either from right to left or from left to right [27]. BERT reads the entire string of words all at once. By considering the surrounding words, this method aids the model in understanding the context of each word. Two unsupervised tasks, NSP and MLM are used to pretrain BERT. In the MLM task, 15% of the tokens in every sequence were randomly masked, and the model was taught to anticipate these masked words. In the NSP task, the input consists of pairs of sentences; 50% of them are connected, and the remaining 60% are not. This algorithm is used to forecast which sentence pairings are connected [15].

The BERT-base-uncased model is the currently-used BERT model. Using MLM, it is pretrained in English. No distinction is noted between the terms "English" and "english" because the model is uncased. The model was trained using batches of 256 and one million steps. For 90% of the steps, the length of the sentence was 128 tokens, while for the remaining 10%, it was 512 tokens [15]. In this study, the BERT has been improved and fine-tuned to better fit the demands of the fake news detection task. The BERT has been improved in this study in the following manner: 1) the model's final layer is unfrozen (only the last layer of the BERT model is allowed to be updated and the first 11 layers are not allowed to be

TABLE III
BEST PARAMETERS THAT WERE ACQUIRED

Splitting criteria	Un-freezed layers	Epochs	Batches
First	1	2	10
Second	1	2	10

updated during the training and have the values of the same weights), followed by its training using news articles to obtain the best accurate representation of the input text and 2) two different splitting criteria are employed; criterion 1 used 70% of the data for training, 15% for validation, and 15% for testing. The second criterion uses 70% for training, while 30% is used for testing. To find the ideal model parameters, the trials are repeated numerous times. It has been discovered that, for both splitting criteria, using two epochs and ten batches, respectively, and unfreezing the final layer is optimum [15]. The best results are displayed in Table III.

V. RESULTS

This study aimed to differentiate between detailed fake news and real news. As was previously indicated in the technique section, an improved BERT model was employed for this purpose. BERT is the model of choice for conducting experiments as it considers the context of the input text. In addition, by initializing the model parameters with the values of all BERT parameters, we can increase the performance without starting from scratch.

As we care about the accuracy of the model to measure how much it is accurate in getting the right prediction for each instance, we have to make sure to measure the model ability to find the positive class and how much it is accurate when classifying it as positive which achieved by using the $F1$ score measurement. $F1$ score is the harmonic mean of the precision and recall and it is the most suitable metric when the data is imbalanced, since the accuracy is not a good one because the number of positive class is much smaller or larger than the negative class. The researchers determined the model's performance using a variety of indicators. All indicators show great performance, indicating that the model is reliable enough to identify bogus news with high accuracy. Both splitting criteria have achieved excellent results after training with the settings listed in Section IV, where the first one achieved the values of 0.27%, 99.94%, 99.94%, 99.94%, and 99.944% for loss, accuracy, precision, recall, and $F1$ score, respectively. The second splitting criterion presented values of 0.31%, 99.96%, 99.96%, 99.97%, and 99.96% for loss, accuracy, precision, recall, and $F1$ score, respectively. Results are shown in Table IV. As can be seen, the results are extremely close to one another. The best $F1$ score and accuracy rate of 99.96% and 99.96%, respectively, were noted in this study, which indicated that the proposed model outperformed the previously-mentioned models. The results indicate that this model has a very low error percentage and a higher accuracy for both the splitting criteria, which makes it very trustworthy and reliable.

TABLE IV
RESULTS NOTED IN THIS STUDY

Splitting criteria	Loss	Accuracy	Precision	Recall	F1 Score
First	**0.27%**	99.94%	99.94%	99.94%	99.94%
Second	0.31%	**99.96%**	**99.96%**	**99.97%**	**99.96%**

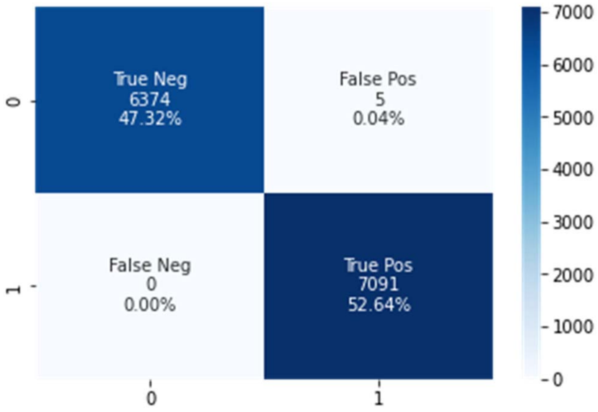

Fig. 4. Confusion matrix on testing data.

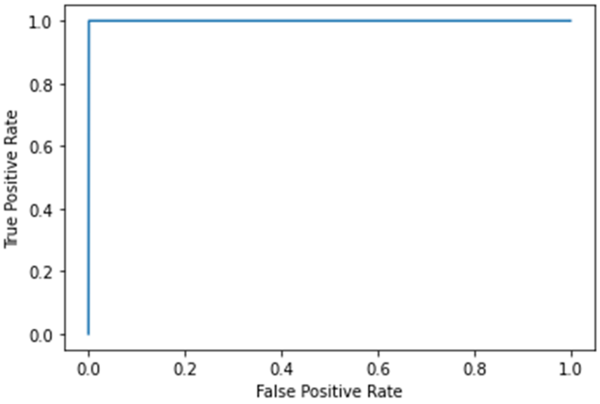

Fig. 5. ROC curve on testing data.

Fig. 4 depicts the confusion matrix for testing data. The false negative value was seen to be 0 in Fig. 4, which indicated that the model could properly classify all the positive values and allocated them to the appropriate class. Five false positives indicate that just five occurrences of the negative class were incorrectly classified as belonging to a positive class by the model. This suggested that only five out of the entire dataset's 44 898 examples were wrongly identified, providing further proof of the model's robustness.

Fig. 5 displays the receiver operating characteristic (ROC) curve which is used to evaluate the performance of binary classification algorithms and provides a graphical representation of the classifier's performance, rather than a single value. ROC curve is constructed by plotting the true positive rate (TPR), or Sensitivity, against the false positive rate (FPR), or (1-Specificity). The true positive rate is the proportion of the instances that were correctly predicted to be positive out

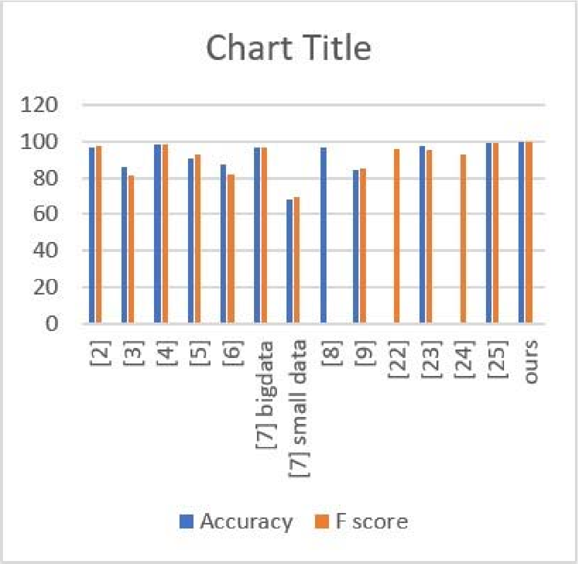

Fig. 6. Comparison of the results noted in this study and those published in earlier studies.

TABLE V
COMPARISON BETWEEN THE RESULTS AND THOSE PRESENTED IN THE LITERATURE

Reference Number	Accuracy	F score
[2]	97.20%	97.60%
[3]	86.00%	81.00%
[4]	98.36%	98.12%
[5]	90.40%	92.80%
[6]	87.42%	82.10%
[7]	96.80% for big data 68.00% for small data	96.80% for big data 69.30% for small data
[8]	96.89%	Not mentioned
[9]	84.40%	85.00%
[22]	Not mentioned	95.78%
[23]	97.50%	95.00%
[24]	Not mentioned	93.30%
[25]	98.80%	98.90%
Ours	**99.96%**	**99.96%**

of all positive instances and is given in the following formula:

$$\text{TPR} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

where TP is the number of true positives and FN is the number of false negatives. The FPR is the proportion of instances that are incorrectly predicted to be positive out of all negative instances and given in the following formula:

$$\text{FPR} = 1 - \text{Specificity} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{2}$$

where FP is the number of false positives and TN is the number of true negatives. The classifier that produces a curve closer to the top-left corner indicates a better performance, whereas the closer the curve comes to the 45° diagonal of the ROC space, the less accurate the test is. As we can see from Fig. 5, the curve is very close to top-left corner which indicates that the classifier has very accurate performance.

Also, the performance of the classifier can be measured by finding the area under the ROC curve (AUC). The higher the AUC score, the better the classifier performs. And from Fig. 5, we can notice that the ROC curve makes it quite evident that the AUC value is $\approx$1, indicating that the model can differentiate between the classes, which further proves the model's robustness.

When compared with the reviewed literature, the proposed model considers the text's temporal degree of information and contextual significance. The experiment is performed using the entire text of the articles, not just the headlines, which is a large dataset. Since the articles included in the dataset describe a variety of subjects and are not exclusive to one particular field, the model will act in the same manner as with different data sets. Table V and Fig. 6 present the comparison of these results to those published in earlier studies. As mentioned before, the accuracy is a very important metric to measure how much the model is accurate in getting the right prediction for each instance. And $F1$ score measures the model ability to find the positive class and how much it is accurate when classifying it as positive. Hence, accuracy and $F1$ score were the chosen metrics for comparison and it seems clear that our results outperform the rest studies.

VI. CONCLUSION

Multiple studies have used deep learning and machine learning to detect fake news on social media. Although deep learning approaches outperform machine learning techniques, they are still not reliable enough because they require large datasets and extensive training to achieve superior results. Pretrained models enable researchers to employ smaller datasets, shorten training times, and obtain a significantly more reliable performance. The pretrained BERT model is used in this study's tests to identify fake news in the dataset that was downloaded from the Kaggle website. The experimental results were better than those shown by other trials, with an accuracy rate and $F1$ score of 99.96% and 99.96%, respectively. To ensure that pretrained models can produce a better outcome, the researchers would conduct further experiments utilizing larger datasets with different languages. However, if two datasets in Arabic and English were combined, the model behavior will be different absolutely. The BERT model used in the study is pretrained on the English dataset, so it will behave well with data from the same domain. Arabic data have different domains so the model will not be able to classify or cluster it in vector space in the right way. As a part of future work, we need to classify Arabic data, the used model preferred to be pretrained on Arabic data to give better performance.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[2] J. Ayoub, X. J. Yang, and F. Zhou, "Combat COVID-19 infodemic using explainable natural language processing models," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102569.

[3] J. C. S. Reis, A. Correia, and F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intell. Syst.*, vol. 34, no. 2, pp. 76–81, Mar./Apr. 2019.

[4] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," *Cogn. Syst. Res.*, vol. 61, pp. 32–44, Jun. 2020.

[5] 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*, 2019, pp. 395–405.

[6] S. B. S. Mugdha, S. M. Ferdous, and A. Fahmin, "Evaluating machine learning algorithms for Bengali fake news detection," in *Proc. 23rd Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2020, pp. 1–6.

[7] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Phys. A, Stat. Mech. Appl.*, vol. 540, Feb. 2020, Art. no. 123174.

[8] L. Singh, "Fake news detection: A comparison between available deep learning techniques in vector space," in *Proc. IEEE 4th Conf. Inf. Commun. Technol. (CICT)*, Dec. 2020, pp. 1–4.

[9] M. B. Yassein, S. Aljawarneh, and Y. Wahsheh, "Hybrid real-time protection system for online social networks," *Found. Sci.*, vol. 25, no. 4, pp. 1095–1124, 2019.

[10] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.

[11] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: A transformer-based approach," *Int. J. Data Sci. Anal.*, vol. 13, pp. 335–362, Jan. 2022.

[12] E. Amer, K.-S. Kwak, and S. El-Sappagh, "Context-based fake news detection model relying on deep learning models," *Electronics*, vol. 11, no. 8, p. 1255, Apr. 2022.

[13] A. Bovet and H. A. Makse, "Influence of fake news in Twitter during the 2016 U.S. Presidential election," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, Dec. 2019.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[16] *Fakenews1*. Accessed: May 25, 2022. [Online]. Available: https://www.kaggle.com and www.kaggle.com/datasets/kruzes1/fakenews1

[17] S. Qin, J. Zhu, J. Qin, W. Wang, and D. Zhao, "Recurrent attentive neural process for sequential data," 2019, *arXiv:1910.09323*.

[18] K. Darwish, W. Magdy, and T. Zanouda, "Improved stance prediction in a user similarity feature space," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 145–148.

[19] T. Elsayed et al., "Overview of the CLEF-2019 CheckThat! Lab: Automatic identification and verification of claims," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Lugano, Switzerland: Springer, 2019, pp. 301–321.

[20] F. Harrag and M. K. Djahli, "Arabic fake news detection: A fact checking based deep learning approach," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 4, pp. 1–34, Jul. 2022.

[21] H. Himdi, G. Weir, F. Assiri, and H. Al-Barhamtoshy, "Arabic fake news detection based on textual analysis," *Arabian J. Sci. Eng.*, vol. 47, pp. 10453–10469, Feb. 2022.

[22] M. S. H. Ameur and H. Aliane, "AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset," *Proc. Comput. Sci.*, vol. 189, pp. 232–241, Jan. 2021.

[23] 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.

[24] 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.

[25] A. B. Nassif, A. Elnagar, O. Elgendy, and Y. Afadar, "Fake news detection in Arabic tweets during the COVID-19 pandemic," *Neural Comput. Appl.*, vol. 12, no. 6, pp. 121–129, Sep. 2021.

[26] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[27] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," OpenAI, USA, Tech. Rep., 2018.