DOI: 10.24996/ijs.2025.66.2.18

ISSN: 0067-2904

Leveraging Arabic BERT for High-Accuracy Fake News Detection

Raed S. Matti*, Suhad A. Yousif
Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

Received: 3/10/2023 Accepted: 26/1/2024 Published: xx

Abstract

Media platforms have become essential for staying informed about events and activities around the globe. While there has been research on identifying news in English, detecting it in Arabic has been relatively overlooked. The unique linguistic characteristics and diverse slang expressions in Arabic have contributed to a scarcity of studies in this area. This research examines the effectiveness of deep learning algorithms in identifying fake news, specifically in the Arabic language. In this study, Global Vectors for Word Representation (GloVe) were used to capture the semantic relationships between words in order to improve the performance of the models. Furthermore, the utilization of Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) algorithms has shown potential in the field of neural networks for text classification purposes. The study also delves into incorporating a trained Arabic BERT (Bidirectional Encoded Representations from Transformers) model, which is widely recognized for its outstanding performance, in various natural language processing tasks. The current research utilizes the Arabic Fake News Dataset (AFND). It is a large, fully annotated dataset of Arabic fake news. The results demonstrated that, among all the investigated algorithms, Arabic BERT achieved accuracy with a score of 0.98 on the dataset. According to the results obtained, LSTM and BiLSTM achieved scores of 0.94 and 0.93, respectively, implementing GloVe word embeddings. This research showcases the effectiveness of the Arabic BERT model, alongside the LSTM and BiLSTM models, in detecting information. It highlights the contribution of BERT in enhancing accuracy when dealing with the identification and mitigation of challenges related to identifying news in Arabic-language contexts.

Keywords: Bidirectional Encoder Representations from Transformers, Deep Learning, Fake News, Natural Language Processing, Word Embedding.

الاستفادة من النموذج العربي لتقنية تمثيلات التشفير ثنائية الاتجاه من المحولات للكشف عن الأخبار المزيفة بدقة عالية

رائد سعد الله عبد الاحد، سهاد عبد الرحمن يوسف
علوم الحاسبات، كلية العلوم، جامعة النهرين، بغداد، العراق

الخلاصة

لقد اصبحت وسائل التواصل الاجتماعي مهمة للاطلاع على الاحداث والنشاطات حول العالم، في حين ان هناك ابحاث مهمة تكشف الاخبار المزيفة باللغة الانكليزية، الا ان اكتشافها باللغة العربية لم يحظى باهتمام كبير

*Email: raedmatti.wk@gmail.com

بسبب الخصائص اللغوية الفريدة والجهود البحثية المحدودة في هذا الصدد. نستنتج من هذه الدراسة مدى اهمية خوارزميات التعلم العميق المختلفة لكشف الاخبار المزيفة باللغة العربية. لقد استعمل الباحثون المتجهات العلمية لتمثيل الكلمات لالتقاط العلاقات الدلالية بين الكلمات ولتعزيز اداء النماذج. بالإضافة الى ذلك، اظهرت تطبيقات الشبكات العميقة مثل نموذج الذاكرة القصيرة والطويلة الأمد ذو الاتجاهين ونموذج الذاكرة القصيرة والطويلة الأمد امكانية كبيرة في تصنيف النصوص عند استعمالها في النماذج العميقة للشبكات العصبية. تشمل الدراسة ايضا النموذج العربي لتقنية تمثيلات التشفير ثنائية الاتجاه من المحولات المدرب مسبقا على اللغة العربية، والذي اظهر اداء مميزا في العديد من مهام معالجة اللغات الطبيعية. تستعمل هذه الدراسة مجموعة بيانات الأخبار المزيفة العربية، فهي مجموعة بيانات كبيرة مفصلة بالكامل للأخبار العربية المزيفة. كانت حصيلة نتائج النموذج العربي لتقنية تمثيلات التشفير ثنائية الاتجاه من المحولات، الذي حصل على درجة 0.98 في مجموعة البيانات، تفوق بالأداء على جميع الخوارزميات الأخرى التي تم فحصها من حيث الدقة. باستعمال تضمين الكلمات، حقق نموذج الذاكرة القصيرة والطويلة الأمد ذو الاتجاهين ونموذج الذاكرة القصيرة والطويلة الأمد درجات دقة بلغت 0.94 و0.93 على التوالي. يبين هذا البحث فعالية النموذج العربي لتقنية تمثيلات التشفير ثنائية الاتجاه من المحولات إلى جانب نموذجي الذاكرة القصيرة والطويلة الأمد ذو الاتجاهين ونموذج الذاكرة القصيرة والطويلة الأمد في كشف المعلومات. ويسلط الضوء على مساهمة النموذج العربي لتقنية تمثيلات التشفير ثنائية الاتجاه من المحولات في تعزيز الدقة عند التعامل مع تحديد وتخفيف التحديات المتعلقة بتحديد الأخبار، في سياقات اللغة العربية.

1. Introduction

The term "fake news," which is also referred to as "misinformation," involves the creation of news articles that mimic information, aiming to deceive people into believing and acting upon them as if they were genuine [1]. Misinformation has become a concern in the field of politics and is impacting our society. With the use of the Internet and social media, sharing content has become easier. Being able to detect and identify such misleading information requires attentiveness and focus [2]. The spread of misinformation brings about risks and consequences as it distorts perceptions among individuals and social groups. Moreover, the airing of fake information within social groups can have significant mental impacts, presenting a possible threat to national security and possibly contributing to violence and other related effects [3]. Detecting news in English has garnered significant interest from researchers who have utilized various techniques such as machine learning (ML) [4] and deep learning (DL) [5]–[8]. However, there has been limited exploration of detecting news in other languages, including Arabic. Numerous research studies have explored the potential of employing machine learning algorithms to detect content in Arabic articles. Classifiers were developed in those investigations utilizing a mix of feature engineering and human feature selection, with different levels of success across a range of datasets [9]. Identifying fake news is a complex task that cannot be resolved solely using natural language processing (NLP). Even for humans, it is nearly impossible to ascertain the authenticity of a piece of writing by merely reading it. Additional fact-checking is essential to confirm the accuracy of news beyond any doubt [10]. This study has uncovered certain obstacles, such as the need for large, annotated datasets and the difficulty of adapting machine learning models to other domains and languages. Despite the advancements in Arabic fake news detection, more study is required, especially concerning various platforms and domains. This study seeks to provide an empirical contribution by conducting a comprehensive analysis of multiple deep learning algorithms to evaluate their effectiveness in detecting fake news within a sizable dataset of Arabic language content. The research contributions have been summarized as follows:

- Comparison of Different DL Algorithms: The study contributes by comprehensively comparing various DL algorithms tailored explicitly for detecting fake news in a large Arabic language dataset.

- Utilization of GloVe Word Embeddings for Improved Accuracy: This study contributes by incorporating GloVe word embeddings, a widely used word representation technique, to enhance the accuracy of the DL models.
- Investigation of Transformer-Based Models, Specifically Arabic BERT: Another noteworthy contribution of this research is the investigation of transformer-based models, with a specific focus on Arabic BERT.

The format of the present research is as follows: The second part of this paper is a literature study on the topic of identifying fake news. In Section 3, an outline of the different ways to spot false or misleading information is given. The methodology utilized in this investigation is described in Section 4. Model evaluations and explanations can be found in Section 5. The outcomes of our proposed methods for detecting fake news are summarized in Section 6.

2. Related Work

Numerous researchers have developed machine learning models for identifying fake news since it has become a severe problem. However, earlier works show that there needs to be more research on spotting fake news in Arabic. The lack of a significant dataset to develop predictive models, the prevalence of slang in spoken and written language, a lack of resources, and insufficient assistance for the Arabic language in libraries are just a few of the challenges and barriers associated with the Arabic language. Tahseen et al. [11] suggest utilizing deep learning techniques to prevent the spread of false news in Arabic-speaking languages. The dataset used in the experiment was called AraNews. 20,300 items have been categorized as false or non-fake in this dataset. The dataset has been reduced by 16,600 articles after preprocessing, of which 8,406 are fake and 8,194 are not fake. The suggested technique is a hybrid deep neural network that integrates convolutional neural networks (CNNs) with LSTM networks. CNNs extract textual properties, whereas LSTM networks collect the temporal relationships between words. According to the AraNews dataset, the hybrid network achieves an accuracy rate of 91.4%, which is higher than the accuracy scores of Text-CNN and LSTM. Text-CNN and LSTM have accuracy scores of 85.9% and 87.8%, respectively [11]. Mohammad et al. [12] worked toward creating a dataset of Arabic clickbait news and implementing the performance of ML models in detecting such news. The authors collected 3235 news samples, out of which 583 were identified as clickbait. They tested ML models, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Stochastic Gradient Descent (SGD), K-Nearest Neighbor (KNN), and Decision Tree (DT). The evaluation of these models on the tested dataset yielded a macro F1-score ranging from 0.18% to the maximum value of 0.81% [12]. In a study conducted by Bilal et al. [13], the authors proposed a method to enhance the accuracy of news article detection. The authors utilized a database called Arabic False News Database (AFND), which consists of 606,912 articles. To optimize the selection of features, they carefully chose 30,000 fake news articles and 30,000 genuine news items. Different classification algorithms, such as (LR), (SVM), (RF), (KNN), (NB), and AraBERT (a transformer model based on BERT), were employed. Among them, AraBERT performed the best with a F1-score of 0.97 [13]. Jude et al. [14] introduced a corpus specifically designed to examine false information in Arabic through the use of textual entailment techniques. This corpus consisted of 4,547 false and true claims, 3,786 pairs of stances, and supporting evidence. The researchers developed two machine-learning models for claim validation and stance prediction. The primary model utilized trained BERT representations and achieved F1-scores of 76.7% for stance prediction and 64.3% for claim verification. On the other hand, the second model based on LSTM achieved an accuracy of 70.6% [14]. Khaled et al. [15] introduced a model architecture to detect news written in Arabic. The focus of this approach primarily revolved around the analysis of content. To achieve this, the authors combined DL and ML techniques. The different DL models, including CNN + BiLSTM,

BiLSTM, LSTM, CNN + LSTM, and CNN, were explored. Interestingly, when it came to ML algorithms, none of the models outperformed other DL models across all datasets examined. However, the study highlighted the BiLSTM model as the most accurate among them, with an accuracy level of 77%. These findings were based on three datasets: dataset 1 consisting of 1,980 tweets, dataset 2 comprising 2,578 tweets, and dataset 3, which merged both datasets, resulting in a comprehensive set of 4,561 tweets [15]. Neural techniques have not been extensively studied in the field of Arabic fake news detection (FND). This research strives to fill this gap by investigating the effectiveness of neural-based approaches in identifying news in Arabic. Additionally, it aims to offer insights for studies by utilizing a comprehensive dataset and achieving a high score of accuracy.

3. Theoretical Background

This research examines models that are widely used for categorization in natural language processing (NLP) tasks, such as Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Bidirectional Encoder Representations from Transformers (BERT). It also investigates the use of GloVe-based word embedding techniques to enhance word representation in a vector space. The research delves deeper into analyzing the effectiveness and enhancements in convergence that can be achieved by utilizing adaptive moment estimation (ADAM) optimization techniques. Through methodologies, the study thoroughly examines these categorization models and their corresponding elements.

3.1. GloVe Word Embedding

The method known as GloVe (Global Vectors for Word Representation) is a word embedding method that has been developed by researchers at Stanford University [16]. GloVe is a model that was trained on a massive dataset using unsupervised training to gain the embedding matrix for the words to know how near the words are to each other, represent the words near or far from each other, and represent these words in digital form. GloVe depends on co-occurrence statistics and probability ratio tables to generate an embedding matrix for a variety of words [16]. For instance, the vectors representing the concepts of "king" and "queen" will be closer together than those representing "car" and "house."

The incorporation of GloVe embeddings has proven beneficial, as these pre-trained models provide a starting point for neural network-based models. Researchers have demonstrated that integrating GloVe embeddings enhances the performance of these models in NLP tasks such as sentiment analysis, language modeling, and machine translation [17].

3.2. LSTM

In 1997, Hochreiter and Schmidhuber introduced a Long Short-Term Memory (LSTM) network [18]. Different from feedforward networks, LSTMs have feedback connections. A typical LSTM module consists of a cell, an input gate, an output gate, and a forget gate. These components work together to control the flow of information inside the cell system. The cell can store values over time intervals, allowing for information to be transmitted within a loop network. In this loop, each network receives data from other networks, performs operations, generates output, and simultaneously sends data to subsequent networks [19]. Equations (1-6) provide the representation of the LSTM cell.

$$f_t = \sigma(W.[h_{(t-1)}, X_t] + b1) \quad (1)$$

$$i_t = \sigma(W.[h_{(t-1)}, X_t] + b_2) \quad (2)$$

$$C1_t = tanh(W.[h_{(t-1)}, X_t] + b_3) \quad (3)$$

$$o_t = tanh(W.[h_{(t-1)}, X_t] + b_4) \quad (4)$$

$$c_t = f_t \times C_{(t-1)} + i_t \times C1_t \quad (5)$$

$$h_t = o_t \times tanh(C_t) \quad (6)$$

The forget gate (f-t) and input gate (i-t) determine whether to retain or include the cell state (C1-t) based on the input and the previous hidden state. The output gate (o-t) regulates the transmission of information from the cell to the state. The dynamic adjustments made by these gates play a role in shaping the evolution of the cell state (c-t), ensuring relevant information is preserved. Using the output gate, the hidden state (h-t) combines the characteristics of the cell state.

3.3. BiLSTM

BiLSTM (Bidirectional Long Short-Term Memory) is a type of recurrent neural network (RNN) that aims to detect long-term connections in sequential input data. It is an improvement over RNNs as it specifically tackles the issue of the vanishing gradient [18]. The functional mechanism of BiLSTM is to feed the inputs into a network, and each of these inputs is an input to forward and backward directions. The sigmoid activation function applies the summation of the outputs for the backward and forward directions for the same input to give the final output of this input, and the process keeps on until the last instance in the input dataset [20].

3.4. BERT

The transformer idea was originally created in recent years. The transformer consists of the encoder and decoder parts. The encoder turns the input sequence into a higher-dimensional space. One of the most important models developed in this field is BERT (Bidirectional Encoder Representations from Transformers) by Google [21]. BERT is a pre-trained model and tokenizer to perform sequence classification on a dataset. The model has been pre-trained on a large corpus of text and is capable of performing multiple NLP tasks, such as BERT, which made an improvement by implementing a masked language model (MLM) during the pre-training step [21]. This method includes masking parts of a text, which encourages the model to predict the words based on the context provided by the masked parts. By utilizing this technique, BERT can grasp the connections between words and variations in context, resulting in improved performance [22]. BERT's ability to process information bidirectionally is also worth mentioning. BERT can take into consideration both the preceding and subsequent contexts of a word during prediction, in contrast to earlier language models, which were unidirectional and could only evaluate text in one way (for example, left-to-right) [23]. Figure 1 illustrates the pre-training and fine-tuning techniques employed by BERT.

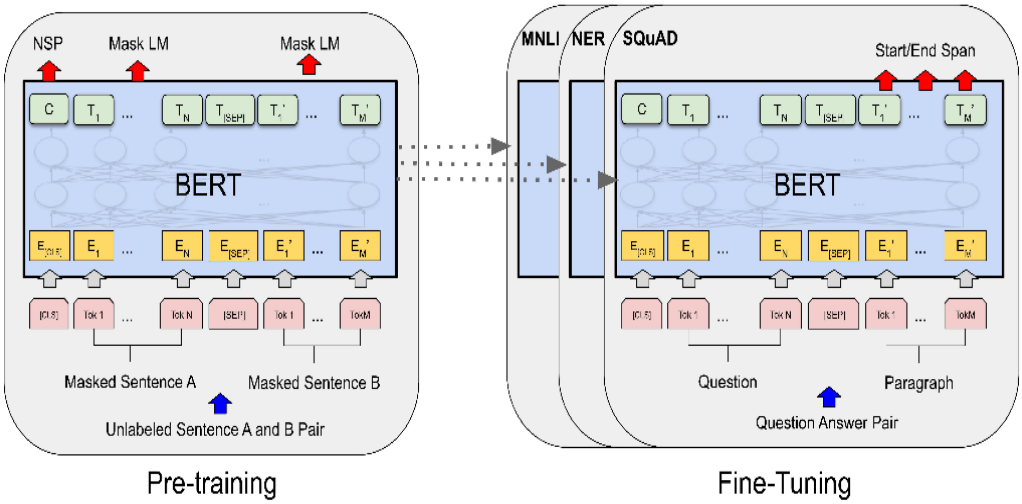

Figure 1: Pre-training and fine-tuning techniques used by BERT [21]

3.5. Optimization Method

Machine learning algorithms that desire to increase the accuracy and speed of training models should incorporate optimization methods. ADAM [24] (Adaptive Moment Estimation) is a popular optimization technique that is regularly utilized in deep learning. Adjusting the learning rate for each parameter depending on the estimates of the first and second moments of the gradients is a concept of improving the adaptive learning rate that assists the model in converging faster [24]. ADAM extends a more functional and dependable optimization strategy by making a combination of the advantages of two popular optimization algorithms, RMSprop and Adagrad.

4. Proposed Methodology

The current research used the Arab Fake News Dataset (AFND) to determine instances of fake news. First, we prepared the dataset by normalizing the text, removing stop words, special characters, and emojis, and performing stemming. After that, we incorporated GloVe word embeddings into both BiLSTM and LSTM models using the preprocessed dataset. To assess the model's performance, we used accuracy to measure the proportion of predictions. We also employed precision, recall, and F1-score as evaluation metrics to provide an assessment of the model's effectiveness. In addition, we used the Arabic BERT model [25] and the Arabic BERT tokenizer tool to evaluate the effectiveness of transformer-based models in detecting fake news in the Arabic language. Figure 2 shows the architecture of the proposed system.

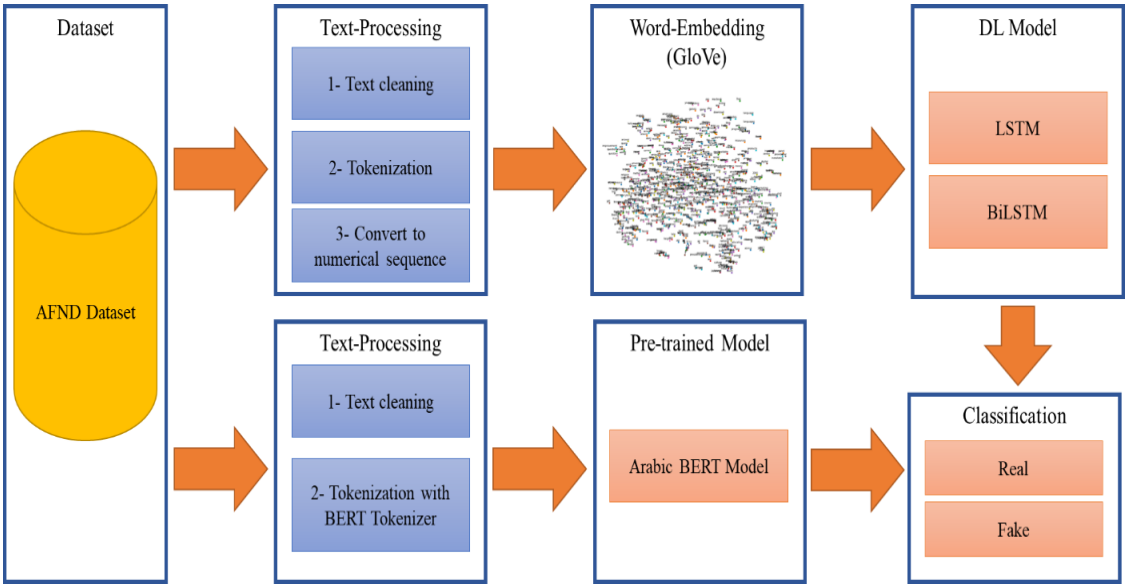

Figure 2: System architecture diagram

4.1. Dataset

In this study, the Arabic fake news dataset (AFND) [26] was utilized as the main source of data in JSON format. We converted the dataset into tables to prepare the data for the experiment. AFND is a large, annotated, and diverse Arabic fake news dataset compiled from publicly available Arabic websites. This dataset comprises 606,912 articles collected in six months from 134 public news websites in 19 Arab countries. The credibility of the news sources was professionally assessed using the Arabic fact-checking tool Misbar, which classified them as undecided, credible, or not credible. For the experiment, we chose a sample from the dataset and sorted it into two identified groups: true news and false news. There were 83,298 recordings in the sample, 42,381 of which were classified as accurate and 40,917 as fake news. We divided the dataset in half, using 80% for training and 20% for testing. Figure 3 shows an overview of the dataset's most frequently used words, broken down by real and fake news sources.

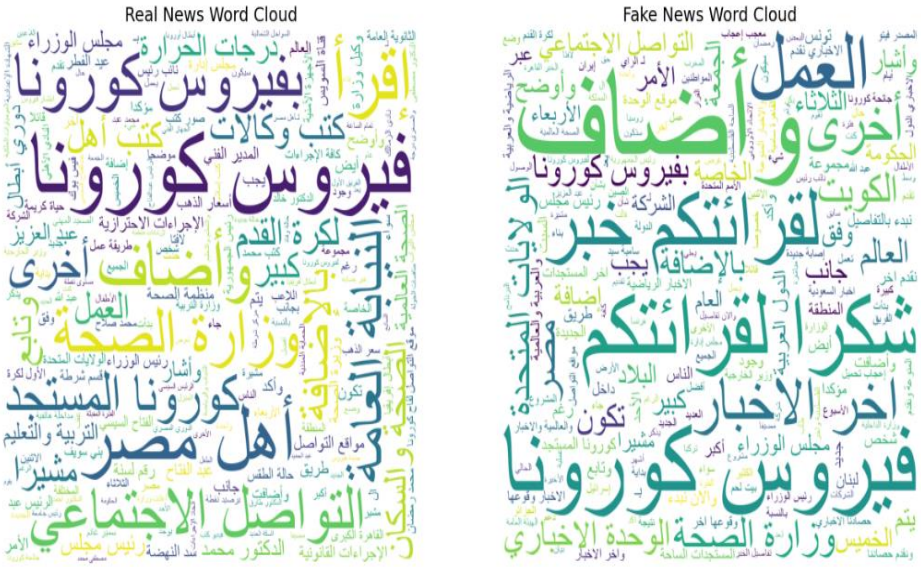

Figure 3: Word cloud of the dataset.

4.2. Preprocessing phase

In this work, we preprocessed the Arabic language dataset to prepare it for analysis. The initial stage combined the news headline with the article and translated the numerals in Arabic to their English equivalents. Then, we deleted any emojis or special characters from the text. We then normalized the Arabic text by deleting diacritics and normalizing the letter forms. We also eliminated Arabic stop words, which are often used but add nothing to the text's meaning. Finally, we applied stemming to the full text with the Tashaphyne: Arabic Light Stemmer [27]. Stemming is a popular technique to reduce words to their simplest form, simplifying analysis and enhancing accuracy. After these preprocessing steps, we saved the cleaned dataset in two parts: the clean text and the stemmed text.

4.3. Tokenization and Padding

Tokenizing the text input and padding the generated numerical sequences to a set length were used to pre-process it. First, tokenization used the Keras library's tokenizer function, with the maximum number of features set at 15,000. Consequently, each word in the dataset was assigned a unique integer index. Following that, the tokenized sequences were padded to a constant length of 250 using the Keras library's pad sequences function to guarantee that they all had the same length and the model could process them. In addition, we used the Arabic BERT tokenizer to tokenize the words and feed them into the Arabic BERT model for fine-tuning. This stage was necessary for the neural network to process the incoming data effectively.

4.4. Proposed Models Architecture

The proposed models utilized the sequential class from the Keras library. The sequential model is a linear stack of layers, and it is possible to add layers to the model using the add method. The model generates a single class label from a series of text inputs (a binary label, in this case, given the use of the sigmoid activation function in the final layer). It used GloVe embeddings as pre-trained word embeddings to provide the model with additional semantic information. We calculated the mean and standard deviation of the embedding vectors and created an embedding matrix using random values from a normal distribution. We inserted pre-

trained words into the matrix based on the index of corresponding words in the tokenizer's vocabulary. The first model (BiLSTM) consists of the following layers:

- An embedding layer maps the input data (sequences of integer tokens representing words) to a continuous vector space. This layer has an input dimension of max_features and an output dimension equal to the number of columns in the embedding matrix. The weights argument specifies the embedding matrix to use.
- Two bidirectional layers are wrapped around an LSTM layer. The bidirectional wrapper allows the processing of the input data in both forward and backward directions, which improves the model's ability to capture contextual dependencies in the data.
- Two dense layers using Rectified Linear Unit (ReLU) activation are fully connected layers. As a result, these layers learn non-linear combinations of the input features.
- The last dense layer produces a binary class label with a single unit and a sigmoid activation function.

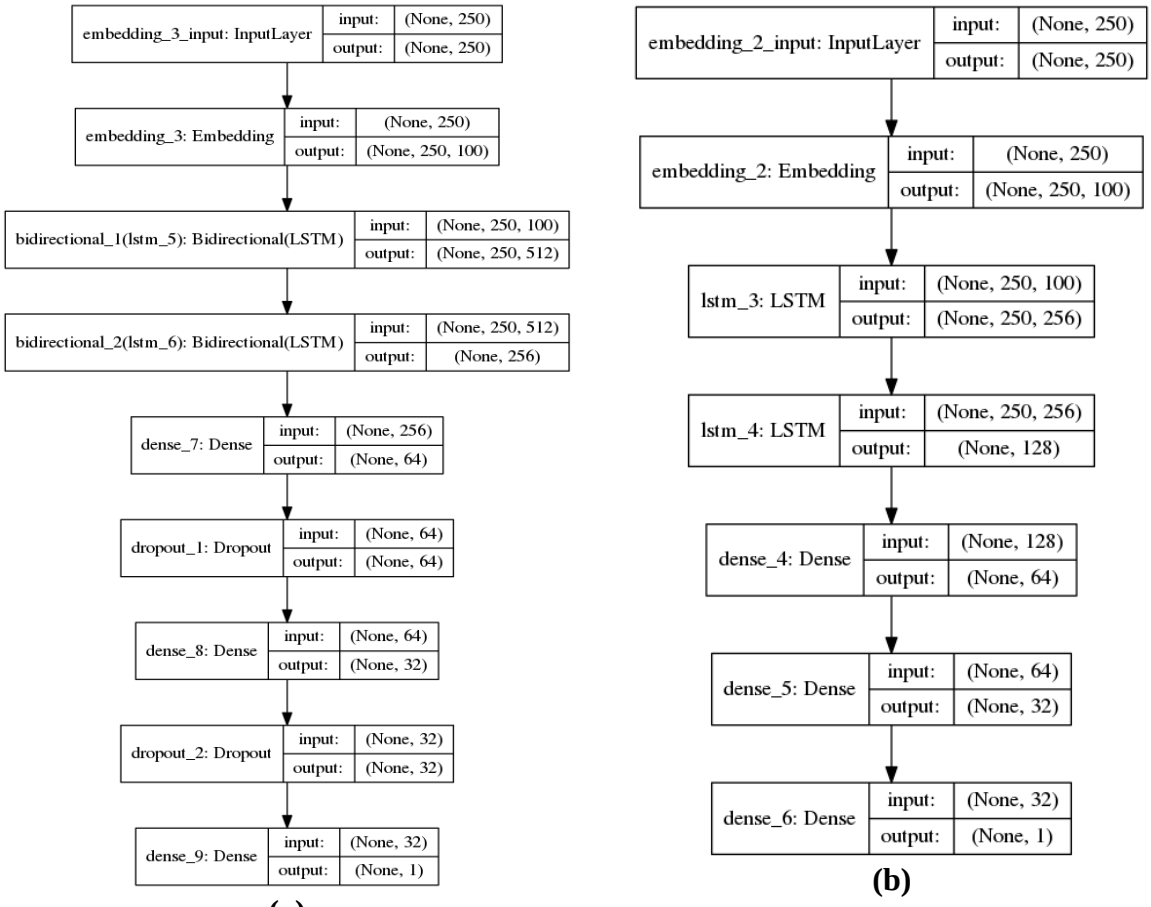

Figure 4: Proposed architecture: (a) BiLSTM and (b) LSTM

The second model (LSTM) consists of:

- Embedding Layer: Converts integer word tokens with a maximum sequence length of 250 into continuous vectors with a word embedding size of 100.
- Two LSTM layers. adept at capturing sequential dependencies and extracting high-level characteristics from the given data.
- Two dense layers: are responsible for acquiring non-linear feature combinations from the output generated by the LSTM layers.

- Output Dense Layer: is responsible for generating binary classification outputs. This is achieved by applying a sigmoid activation function.

Figure 4 (a) illustrates the architecture of the BiLSTM model, whereas Figure 4 (b) represents the architecture of the LSTM model.

The models were optimized utilizing the Adam optimizer, with a learning rate of 0.001. The binary cross-entropy loss function was utilized as the training objective. The evaluation was conducted using the accuracy metric. The training process consisted of 12 epochs, with a batch size of 256.

On the other hand, we used the Arabic BERT model with the same settings and hyperparameters but with a smaller number of epochs, four. These settings allowed the model to learn and extract more complicated and detailed features from the Arabic text data, contributing to its performance in detecting fake news articles.

4.5. Evaluation Metrics

In the metric evaluation phase of the DL project, we assess the model's performance using the accuracy metric. Accuracy is the proportion of correct predictions the model makes. During this phase, we calculate the ratio of predictions (both true positive (TP) and true negative (TN)) to the total number of predictions [28]. Also consider the number of false positive (FP) and false negative (FN) predictions made by the models, which are instances where their predictions differ from the actual class. For example, the (TP) stands for real news, whereas the (TN) refers to fake news. (FP), conversely, indicates the amount of false news determined to be genuine. Moreover, (FN) is the number of times real news has been misidentified as false news [29].

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{all samples}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{9}$$

$$\text{F1} - \text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{10}$$

5. Results and Discussion

We created and ran the deep learning models in the Kaggle notebook. Kaggle's GPU P100 accelerator is intended to run TensorFlow. In this part, we provide the findings of determining the accuracy of each model stated in this study. The equations (7–10) were used to calculate precision, F1-score, and recall. The objective of this research is to improve the accuracy of identifying false news by identifying important characteristics via the use of GloVe word embeddings and the Arabic BERT tokenizer. First, data preprocessing eliminates noise from the text before analyzing the news. The stemming technique then transforms the words into their base forms. The next stage involves converting words into features that eventually make their way into the model for perpetration. Finally, it builds models and trains them using the dataset. The study’s findings include categorizing news into two categories: true or false. The experiment operated on two files, the first on the cleaned data only and the second on the clean data with stemming. Our findings indicate that GloVe performs more effectively when applied to unaltered data without the use of stemming techniques. GloVe utilizes the distributional semantics of words in a corpus, allocating vectors to words based on their probability of co-occurrence. Stemming has the capacity to modify the original context and grammatical structure of words, which may have an influence on the accurate depiction of relationships between words. On the other hand, Arabic BERT with the BERT tokenizer had the highest accuracy.

Figure 5 (a) shows an Arabic BERT model's training, validation loss, and accuracy over four epochs. The training loss and accuracy are represented by solid lines, whereas the validation loss and accuracy are represented by dashed lines. Following this, Figures 5 (b) and (c) illustrate the progression of training and validation loss and accuracy throughout 12 epochs for both the BiLSTM and LSTM models.

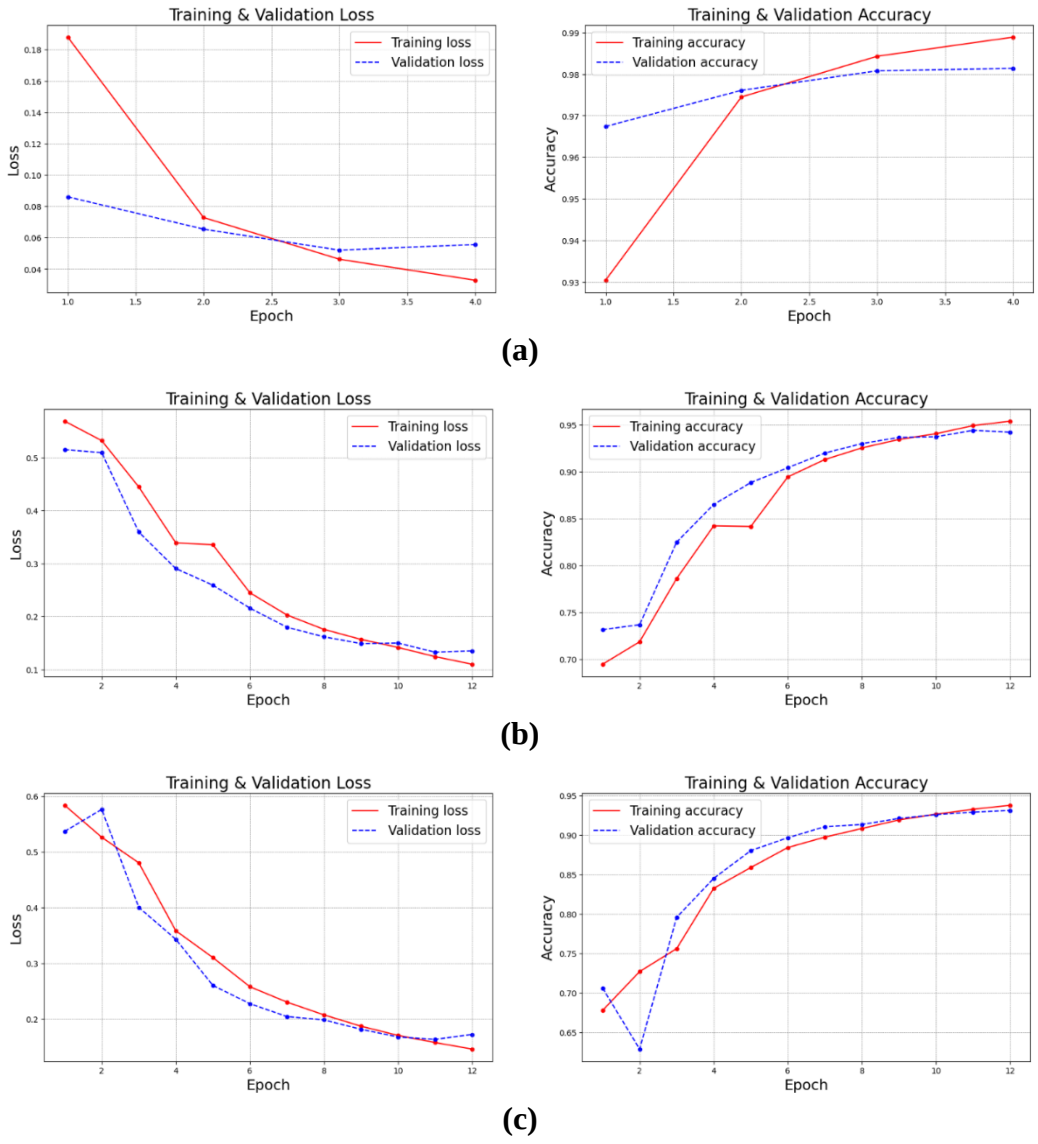

Figure 5: Loss and accuracy of models: (a) Arabic BERT, (b) BiLSTM, and (c) LSTM

The models were trained with the Adam optimization technique to detect bogus news. Figures 4 (b) and (c) illustrate the noticeable convergence pattern seen in the model. Throughout the sequence of 12 epochs, both the training and validation accuracy continually improved, while the loss displayed a steady decline. This implies that the model has acquired the ability to differentiate between fabricated news and authentic news.

The confusion matrix shows that the Arabic BERT correctly identified many articles as either true or false news. Based on the true positive and negative values, the model has effectively identified a total of 8,113 articles containing false news and 8,281 articles containing accurate news. On the other hand, the model misidentified 120 articles of real news as fake news and 146 articles of false news as genuine, as shown by the false positives and false

negatives. Figures 6 (a), (b), and (c) display the confusion matrix for the models when evaluated on the clean dataset.

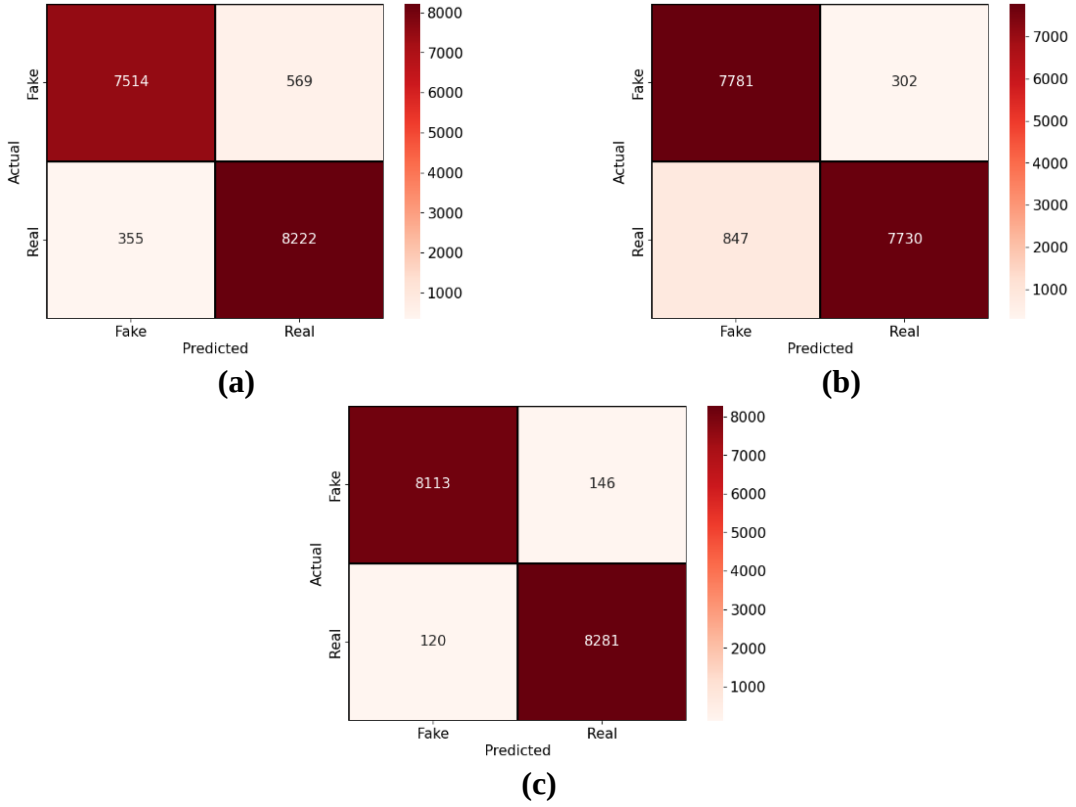

Figure 6: Confusion matrix of models: (a) BiLSTM, (b) LSTM, and (c) Arabic BERT

Tables 1 and 2 display the classification results of the models on the dataset with two different versions: the first version without stemming the text and the second one after applying the stemming method. The results show that Arabic BERT surpasses sequential models BiLSTM and LSTM in many performance parameters. Arabic BERT outperforms BiLSTM (94%) and LSTM (93%), with an accuracy of 98% without stemming. Arabic BERT also has good precision (99%) and recall (99%), allowing it to classify positive instances with few false positives and negatives. The F1-score highlights Arabic BERT's balance. With 97% accuracy after stemming (Table 2), Arabic BERT remains superior to linguistic variations. At 98%, the model's accuracy, recall, and F1-score illustrate its consistency and dependability with stemmed data. While all models perform less well with stemming.

Table 1: Classification report of the models.

	BiLSTM	**LSTM**	**Arabic BERT**
Accuracy	94%	93%	98%
Precision	95%	93%	99%
Recall	94%	93%	99%
F1-score	94%	93%	99%

Tables 1 and 2 demonstrate that our study exceeded previous research efforts in detecting fake news. In particular, in [11], the LSTM model shows a notable accuracy of 87.8%. However, our LSTM model achieved a higher accuracy rate of 93%. In addition, the author in [15] obtained the most remarkable accuracy of 77%, whereas our BiLSTM model obtained a superior accuracy of 94%.

Table 2: Classification report of the models with stem

	BiLSTM	LSTM	Arabic BERT
Accuracy	92%	91%	97%
Precision	92%	92%	98%
Recall	92%	92%	98%
F1-score	92%	92%	98%

In contrast, it is crucial to ensure consistency in conducting a comparative analysis with previous research. This can be achieved by utilizing the same benchmark dataset, AFND, along with the evaluation metric, specifically the F1-score. Significantly, a comprehensive investigation conducted by [13] encompassed the use of many classification methods. The results of this inquiry revealed an F1-score that ranged from 86% to a maximum performance of 97%. In the context of our study, we conducted research using a larger sample of the dataset, consisting of 83,298 records. The results of our study revealed a substantial enhancement in accuracy, with a maximum achievement of 98% and an F1-score of 99%. This highlights the significant progress made in the field of fake news detection. Moreover, it is crucial to emphasize that our deep learning methodology not only attained higher levels of accuracy but also exhibited significant efficiency in terms of training duration in comparison to traditional machine learning models. The average length of each epoch is approximately 95 seconds, thereby demonstrating the usefulness of deep learning concerning both accuracy and computational efficiency when applied to tasks involving the detection of fake news.

6. Conclusion

This study aimed to predict fake news articles using deep learning models. Our proposed model applied GloVe word embedding to measure the relationship between the news article title and body and assess the news's credibility based on co-occurrence probabilities. Our results showed that both the BiLSTM and LSTM models had high accuracy. However, the Arabic BERT transformer model, both with and without a stem, significantly outperformed the LSTM and BiLSTM models. The findings demonstrate the efficacy of the Arabic BERT transformer model in the prediction of fabricated news, suggesting its potential superiority over the LSTM and BiLSTM models for this particular task. However, BiLSTM outperformed LSTM in extracting local and position-invariant features. Additionally, our results indicated that BiLSTM was significantly more effective on the clean dataset without applying the stemming method than unidirectional models. Finally, although our study focused on Arabic news, additional investigation is required to comprehensively comprehend the capabilities of deep learning models, particularly the Arabic BERT transformer word embeddings, in evaluating the automatic credibility analysis of news across different languages and domains.

References

[1] F. A. Mukhaini, S. A. Abdoulie, A. A. Kharuosi, A. E. Ahmad, and M. Aldwairi, "FALSE: Fake News Automatic and Lightweight Solution," *arXiv*, Aug. 2022, Available: https://doi.org/10.48550/arXiv.2208.07686.

[2] R. Jehad and S. A. Yousif, "Fake News Classification Using Random Forest and Decision Tree (J48)," *Al-Nahrain Journal of Science*, vol. 23, no. 4, pp. 49–55, Dec. 2020, Doi: 10.22401/ANJS.23.4.09.

[3] M. Fayaz, A. Khan, M. Bilal, and S. U. Khan, "Machine learning for fake news classification with optimal feature selection," *Soft comput*, vol. 26, no. 16, pp. 7763–7771, Jan. 2022, Doi: 10.1007/s00500-022-06773-x.

[4] S. Senhadji and R. A. S. Ahmed, "Fake news detection using naïve Bayes and long short term memory algorithms," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 746–752, Jun. 2022, Doi: 10.11591/ijai.v11.i2.pp746-752.

[5] I. K. Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, "Detection of fake news using deep learning CNN–RNN based methods," *ICT Express*, vol. 8, no. 3, pp. 396–408, Sep. 2022, Doi: 10.1016/j.icte.2021.10.003.

[6] S. Deepak and B. Chitturi, "Deep neural approach to Fake-News identification," *Procedia Computer Science*, vol. 167, pp. 2236–2243, 2020, Doi: 10.1016/j.procs.2020.03.276.

[7] P. Bahad, P. Saxena, and R. Kamal, "Fake News Detection using Bi-directional LSTM-Recurrent Neural Network," *Procedia Computer Science*, vol. 165, pp. 74–82, 2019, Doi: 10.1016/j.procs.2020.01.072.

[8] R. Jehad and S. A. Yousif, "Classification of fake news using multi-layer perceptron," in *AIP Conf Proc*, vol. 2334, no. 1, pp. 070004–070009, Mar. 2021, Doi: 10.1063/5.0042264.

[9] T. A. Wotaifi and B. N. Dhannoon, "Improving Prediction of Arabic Fake News Using Fuzzy Logic and Modified Random Forest Model," *Karbala International Journal of Modern Science*, vol. 8, no. 3, pp. 477–485, Aug. 2022, Doi: 10.33640/2405-609X.3241.

[10] B. Upadhayay and V. Behzadan, "Sentimental LIAR: Extended Corpus and Deep Learning Models for Fake Claim Classification," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 1-6, Arlington, VA, USA, 2020, Doi: 10.1109/ISI49825.2020.9280528.

[11] T. A. Wotaifi and B. N. Dhannoon, "An Effective Hybrid Deep Neural Network for Arabic Fake News Detection," *Baghdad Sci. J*, vol. 20, no. 4, pp. 1392-1401, Aug. 2023, Doi: 10.21123/bsj.2023.7427.

[12] M. A. Bsoul, A. Qusef, and S. Abu-Soud, "Building an Optimal Dataset for Arabic Fake News Detection," *Procedia Computer Science*, vol. 201, pp. 665–672, 2022, Doi: 10.1016/j.procs.2022.03.088.

[13] B. Hawashin, A. Althunibat, T. Kanan, S. AlZu'bi, and Y. Sharrab, "Improving Arabic Fake News Detection Using Optimized Feature Selection," in *2023 International Conference on Information Technology (ICIT)*, pp. 690–694, Amman, Jordan, Aug. 2023, Doi: 10.1109/ICIT58056.2023.10225974.

[14] J. Khouja, "Stance Prediction and Claim Verification: An Arabic Perspective," *arXiv*, May. 2020, Available: http://arxiv.org/abs/2005.10410.

[15] K. M. Fouad, S. F. Sabbeh, and W. Medhat, "Arabic fake news detection using deep learning," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 3647–3665, 2022, Doi: 10.32604/cmc.2022.021449.

[16] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, 2014, Doi: 10.3115/v1/D14-1162.

[17] X. Ma and E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1064–1074, Berlin, Germany, Aug. 2016, Doi: 10.18653/v1/P16-1101.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, Doi: 10.1162/neco.1997.9.8.1735.

[19] N. A. K. Hussein and B. Al-Sarray, "Deep Learning and Machine Learning via a Genetic Algorithm to Classify Breast Cancer DNA Data," *Iraqi Journal of Science*, vol. 63, no. 7, pp. 3153–3168, Jul. 2022, Doi: 10.24996/ijs.2022.63.7.36.

[20] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, Jul. 2005, Doi: https://doi.org/10.1016/j.neunet.2005.06.042.

[21] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, Oct. 2018, Available: http://arxiv.org/abs/1810.04805.

[22] M. Peters *et al.*, "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, New Orleans, Louisiana, 2018, Doi: 10.18653/v1/N18-1202.

[23] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv*, Jul. 2019, Available: https://doi.org/10.48550/arXiv.1907.11692.

[24] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv*, Dec. 2014, Available: https://doi.org/10.48550/arXiv.1412.6980.

[25] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in social media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 2054–2059, Barcelona, 2020, Doi: 10.18653/v1/2020.semeval-1.271.

[26] A. Khalil, M. Jarrah, M. Aldwairi, and M. Jaradat, "AFND: Arabic fake news dataset for the detection and classification of articles credibility," *Data Brief*, vol. 42, pp. 108141–108148, Jun. 2022, Doi: https://doi.org/10.1016/j.dib.2022.108141.

[27] T. Zerrouki, "Tashaphyne, Arabic light stemmer." 2012, Available: https://pypi.python.org/pypi/Tashaphyne/0.2.

[28] S. A. Yousif, V. W. Samawi, and N. M. G. Al-Saidi, "Automatic Machine Learning Classification Algorithms for Stability Detection of Smart Grid," in *2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI)*, pp. 34–39, Fuzhou, China, 2022, Doi: 10.1109/BDAI56143.2022.9862710.

[29] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv*, Oct. 2020, Available: https://doi.org/10.48550/arXiv.2010.16061.