Human-Centric Intelligent Systems (2024) 4:136–146
https://doi.org/10.1007/s44230-023-00055-x

RESEARCH ARTICLE

Hybrid Neural Network Models for Detecting Fake News Articles

Ashwaq Khalil[2] · **Moath Jarrah**[1,2] · **Monther Aldwairi**[2]

Received: 5 July 2023 / Accepted: 22 November 2023 / Published online: 16 December 2023

Abstract
The prevalence of world-wide access to the Internet has come at a cost. A lot of misleading information is posted on public news websites and social media. Many news writers and organizations manipulate their posted data to propagate false information that target different societies and in different languages. Accurate and timely detection of false news is made possible in large part using machine learning-based technologies. This paper targets the problem of detecting fake news in Arabic language using machine learning models. A hybrid model of two deep neural networks is used to classify Arabic news articles in order to detect fake articles. The two types of neural networks are convolutional and bi-directional long-short term memory. Robust features are extracted using two different word vectors and a complex model of a convolutional neural network. Moreover, a set of auxiliary output layers are used to enhance the model accuracy. Multi-class classification is achieved via modifying the primary output layer. Results show an accuracy of 88% and 78% for binary classification and multi-class classification, respectively.

Keywords Fake news · Neural networks · Multi-class classification · Binary classification

1 Introduction

News target different people who are interested in specific events, topics, or facts [1]. Fake news is defined as unverified and manipulated information that propagates to misguide newsreaders, in order to create incorrect awareness, earn money, or achieve political objectives [1–7]. False information is manipulated by many parties such as individuals, groups, social bots, and news organizations. Moreover, the development of social bots for different platforms of social media has facilitated the dissemination of fake news easily and rapidly [5, 7–9].

False information affects individuals, businesses, governments, and democracy, and it has been shown in recent years that it may cause calamities in societies. It results in a negative impact on journalism, society, economy, political insecurity, elections, and public judgment [1, 4–8, 10–12]. For instance, fake news related to the Corona pandemic affected the safety, physical and mental health of the public [13]. Consequently, the term fake news detection has been formulated recently, which refers to the detection of deceptive news articles that targets people in order to affect their ideas about the topic of interest [2].

In contrast to the conventional media, a huge number of news events and articles propagate rapidly among newsreaders through social media such as Twitter and Meta (i.e. Facebook) and the Internet [1, 6, 7, 12, 14, 15]. Additionally, fake news is distributed as private messages through applications such as WhatsApp. The Internet facilitates the access to read news from anywhere and also to send it rapidly with minimal effort. In addition, a large number of online news sources are gray and make it difficult to distinguish between fake and real news [2, 5–7, 9]. Consequently, different types of fake news have been extended beyond fake and real such as rumor, misinformation, or disinformation [2, 7]. Rumor news is created and propagated in social media such as Meta (i.e. Facebook), while disinformation news is particularly created and propagated to misguide the public.

✉ Moath Jarrah
mhjarrah@eiu.edu

Ashwaq Khalil
moqashwaq@gmail.com

Monther Aldwairi
munzer@just.edu.jo

[1] School of Technology, Eastern Illinois University, Charleston, IL, USA

[2] Computer and Information Technology, Jordan University of Science and Technology, Irbid 22110, Jordan

Misinformation news or false information is sometimes introduced within legitimate news by mistakes [5].

Automatic fake news detection has a significant importance, because manual detection by expert journalists is inconvenient, costly, time-consuming, and cannot handle the large volume of news in today's big data era. Thus, machine learning (ML) and news datasets are needed to identify fake news automatically [2, 5–7, 14, 16]. Nevertheless, automatic fake news detection has a potential challenge, where ML models require a large number of annotated articles that could be suffering from human bias [1, 6, 12].

Regardless of the language, automatic fake news detection is a hot research problem all around the world [8]. There is an observed lack of research in detecting Arabic fake news compared with English fake news and other languages [4, 17]. Furthermore, most available Arabic datasets were collected for different goals such as categorical classification or named entity recognition. For instance, the ArCAR [18], SATCDM [19, 20] have used the SANAD dataset [21] to classify the Arabic news articles based on the news topic such as sport or politics without considering the problem of fake news. The work in [22] used Arabic news articles for name entity recognition.

In this paper, an automatic Arabic fake news detection model is proposed. The model outperforms the work in [23] in terms of performance. In the proposed model, a hybrid neural network model [5] is improved and enhanced by extracting more robust features that allow the model to discriminate between various classes. Specifically, the contributions of this paper are:

1. Two 300-dimensional word vectors representation are generated and fed to two embedded layers (GloVe and FastText layers).
2. Expanding a one-dimensional convolution layer into three two-dimensional layers to extract robust features. The ELU-gate unit [24] is used to decide which features are needed for activation and which ones should preserve their linear property.
3. Bidirectional long-short term memory (Bi-LSTM) is used for the learning of the order of features dependencies in both directions, where two different activation functions are used.
4. A set of auxiliary outputs is used to increase the suggested model's accuracy and the primary output layer is modified to provide a multi-class classification solution.

We aim to integrate the proposed fake news detection model within Internet browsers, where users can be warned and alarmed of the possibility of fake articles on-the-fly. The rest of the paper is arranged as follows. Relevant research on false news detection techniques for Arabic and non-Arabic languages is covered in Sect. 2. A background on vector representations and deep neural networks (DNN) is provided in Sect. 3. The methodology, which includes the dataset, reference model, and the proposed model, is described in Sect. 4. Also, the model architecture is presented in Sect. 4. The evaluation and results of the experiments are presented in Sect. 5 for the multi-class classification and binary detection problems. The paper is finally concluded with some closing observations in Sect. 6.

2 Related Work

Many researchers have focused on using ML models for the detection of fake, rumors, misinformation, or disinformation news propagated through the Internet [14]. Our research group earlier focus was concerned with a simple tool for the detection of Clickbaits and false news in social media sites [25]. Moreover, we have developed a lightweight solution to visualize fake news datasets, where classification, clustering, plots, and correlation were used to analyze the dataset [26]. More recently, we have shifted the focus to Arabic fake news detection, where we collected a dataset and made it publicly available [4, 17]. An overview of the models that have been proposed for the identification of fake news written in Arabic and non-Arabic languages is given in this section.

2.1 Arabic Fake News and Tweets Detection

The Researchers in [1, 4, 15] used ML to detect Arabic fake news using custom features like content- and user-based elements. Moreover, Alzanin and Azmi [1] utilized topic-based and tweet-based features in unsupervised and semi-supervised models to identify Arabic tweets as rumors or not. The authors of [15] combined topic-based and user-based features. Then, they used content verifiability and users' responses polarity to enhance the performance. Moreover, Johnson et al. [4] used sentiment analysis for Arabic tweets to help the detection accuracy.

Other researchers introduced a word-embedding technique in ML models such as the work in [3] and [11]. The authors of [11] used cross-lingual embedding to train English claims. Then they used Arabic claims to evaluate the trained model. On the other hand, the authors of [3] used the n-gram, char-gram, and term frequency-inverse document frequency (TF-IDF) to calculate a score between headline and the corresponding content. Furthermore, some Arabic researchers in different studies used conventional ML methods, such as Naive Bayes, to evaluate their collected datasets [14, 15, 27]. Others used a transformer-based language approach for Arabic stance detection that consists of true and false claims [10, 16]. The neural-based models played a significant role in some studies for Arabic fake news detection [2, 28]. Transformer-based language models and

DNN models were compared in terms of performance by the authors of [2], where the transformer-based language model, called AraBERT v02 [29], has achieved a better performance. The authors of [28] proposed a deep co-learning approach based on a semi-supervised model that uses a combined two convolutional neural networks (CNN) to estimate the Arabic weblogs. The first CNN branch uses the continuous bag-of-words (BOW) model, while the second branch uses a character-level embedding.

The work in [23] used different ML and deep learning models for the binary and multi-class classification tasks in detecting Arabic fake news. The authors evaluated 8 different models which are: capsule networks, deep double Q-learning, deep pyramid CNN, information distilled LSTM, support vector classification, linear SVC, K-Means, and Bayesian gaussian mixture. The dataset that was used is Arabic Fake News Dataset (AFND) [17]. Studies have demonstrated that deep learning techniques outperformed conventional ML models. The capsule networks model has achieved the best results in terms of accuracy for both binary and multi-class classification tasks. Additionally, the authors noted that the trained models had issues with both underfitting and overfitting, indicating that the AFND dataset is difficult and noisy.

2.2 Non-Arabic Fake News Detection

The authors of [8] used different datasets that contain news from three different languages and various features for comparison, which are: CBOW, skip-gram, document-class distance (DCDistance), and 14 sets of textual features. In addition, the authors employed support vector machines, k-nearest neighbors algorithm, gaussian Naive Bayes, and random forest as their four main traditional ML models for training and testing. The authors of [13] used two approaches for the detection of COVID-19 fake news pandemic. The first approach consists of five pre-trained transformer-based language methods, and the second is a mathematically clean training sample.

The global vectors for word representation (GloVe) [30, 31] have been applied in two methods to improve the models' performance. First, the researchers in [12] used the GloVe model to generate a word representation vectors, while the researchers in [5, 6] used the pre-trained GloVe embedding, which is provided by Stanford NLP team [31]. Moreover, the GloVe vectors are used along with Bi-LSTM for the detection of English fake news [5, 6, 12]. The proposed model in [6] classifies news articles into fake or real news. The proposed model in [12] combines the news articles with live features such as the details of the news authors.

The hybrid deep neural network model [5] is a binary classification model that uses both CNN and LSTM to distinguish between bogus and true news. The length of the input sequences is 300 tokens. The 100-dimensional GloVe representation maps input sequences to the corresponding vectors in the embedding layers. The features are then extracted using a one-dimensional max-pooling layer and a one-dimensional convolution layer. The number and size of the kernels are 128 and 5, respectively. The extracted features (48-dimensional vector) are fed into the LSTM layer to learn long-term dependencies. Sigmoid activation function is used in the output layer, which is a fully connected layer, to shrink the output to 1 and decide whether the article is false or true. The loss function in this case is binary cross-entropy. However, the baseline model is for binary classification only. To compare our proposed model with the baseline for the multi-class classification, we have therefore modified the output layer. The fully connected layer is modified to classify news articles into credible, not-credible, and undecided. The loss function is the category cross-entropy.

Other researchers utilized ML with different systems to identify and detect fake news [7, 9]. The authors of [7] integrated ML models in the Chrome browser to identify the fake news posts on Meta (i.e., Facebook). They improved the effectiveness of their strategy by utilizing both content- and user-based features. The FaNDeR [9] utilized a question-answering system with a CNN model to assess the reliability of the news media by classifying their news into false, true, or neutral.

3 Background

3.1 Word Vector Representation

One of the critical stages in the natural language processing (NLP) and text classification task is generating words vectors representation, where a model is used to convert texts to real-valued vectors [5]. Some of these models are TF-IDF models and the BOW. The BOW model counts unique terms within the text, while the TF-IDF model counts the common and rare terms [19]. However, these models have many limitations, such as disregarding words arrangements and the semantic of the text. Also, vector representations have large dimensions [8, 12, 18, 19]. CNN models are used to handle such limitations, where two types of models are used, which are: feature selection and word-embedding. The feature selection models reduce the dimension of the vectors by selecting subset features. The word-embedding models extract the syntactic and semantic features by extracting statistical relations between two vectors [8]. Consequently, fixed length of sequence integers are generated for text classification tasks [5, 6, 8, 19]. The GloVe representation is the most pre-trained word-embedding model used in text classifications studies [5, 6, 19]. GloVe is an unsupervised learning algorithm that converts each word into a value in a high

dimensional vector. Hence, similar words result in values stored in the same location. However, the word-embedding is sensitive to the language and domain of the datasets [5]. Thus, different models and pre-trained vectors were generated for different languages such as FastText [32].

3.2 DNN

DNN has widely been used in different ML applications for it is high-performance and adaptivity. Companies that rely on artificial intelligence (AI) for their business process utilize deep learning to perform the work artificially and automatically [7, 33]. One of the most interesting characteristics of the DNN is the generalization, where the same architecture is used for different datasets and in various applications [33]. DNN has different approaches that are widely used in NLP such as CNN, recursive neural network (RvNN), and recurrent neural network (RNN) [7, 33]. Next, we will discuss CNN and Bi-LSTM, which is an improved version of RNN.

3.2.1 CNN

It is based on the multilayer perceptron (MLP), which consists of connected neurons to extract features and a fully connected MLP to make decisions. It was built for two-dimensional inputs such as images. The CNN inputs in the NLP tasks are one-dimensional (text), where features are extracted using a one-dimensional convolution layer [5, 6, 19]. In the forward process, several fixed size filters are convoluted over the input data to extract abstract features by multiplying their weights with the data [5]. The weights are updated in the backward process by utilizing the differences between the target and predicted values.

3.2.2 Bi-LSTM

The RNN model learns short sequential data such as short sentences. It utilizes previous and current words of input sequences to recognize the entire meaning of the input text. Moreover, LSTM is an enhanced version of the RNN that is designed in order to learn large inputs such as long articles [5–7, 12]. The four gates that make up an LSTM are the input gate, output gate, input modulation gate, and forget gate. These gates correspond to four different parameters. The output of previous hidden states along with the local inputs are used for training. The four values of the gates are computed to decide which information is stored, read, ignored, and written [5, 6]. In addition, the Bi-LSTM is an improved version of the LSTM to learn and extract features in both directions at the same time, which is referred to by forward (past to future) and backward (future to past) [6].

Symbol/shape/ URL	Corresponding Arabic word
Hashtags	هاشتاغ
Emojis	ايموجي
Link or URL	لينك

Fig. 1 Arabic words for hashtag, emoji, and web links

4 Methodology

4.1 Dataset and Pre-processing

AFND dataset consists of about 607,000 articles that were collected from 134 public Arabic news websites [17, 23]. A weak labeling approach was used to annotate the articles as not-credible, undecided, and credible. Articles that contain less than 120 words before cleaning and 100 words after cleaning were ignored. To achieve a balanced dataset, public news websites that contain 4280 Arabic articles or more were selected. Then 4280 news articles were selected randomly from each website. Afterwards, eight websites were randomly selected for each label. Finally, 90% of the selected articles were used for training and 10% were used for testing. As a result, the multi-class classification problem has total numbers of 8988 and 72,786 articles for testing and training, respectively. In addition, the binary classification problem has 6420 and 51,990 articles for testing and training, respectively. The labels of the binary classification problem are credible (real) and not-credible (fake), while the multi-class classification labels are not-credible, credible, and undecided. The Arabic text is cleaned by eliminating non-Arabic words, stop words, white spaces, and punctuation marks [2, 20, 23]. The Arabic terms are normalized to reduce the model input size using the Tashaphyne library.[1] Moreover, many hashtags, emojis, and website links were observed in the text and were replaced by their corresponding Arabic words as shown in Fig. 1 (presented as a table).

5 Reference Model

The hybrid DNN model is a binary classification model that uses both CNN and LSTM to categorize news into fake or true [5]. The length of the input sequences is 300 tokens. The GloVe representation maps are 100-dimensional input sequences to the corresponding vectors in the embedding layers. Then, a one-dimensional convolution layer and a one-dimensional max-pooling layer are used to extract features.

[1] https://pypi.org/project/Tashaphyne/.

The number of kernels is 128 and the size is 5. The LSTM layer receives 48-dimensional extracted features as input.

A fully connected layer is used for the output layer. In addition, a Sigmoid activation function is utilized to classify an article into true or fake via the use of a binary cross-entropy function. The reference model capability is for binary classification only. Therefore, we modified the output layer to achieve a multi-class classification capability and compare the reference model with the proposed model. The fully connected layer classifies news articles into not-credible, credible, and undecided. The loss function is the category cross-entropy.

5.1 Proposed Model

The proposed model is an enhanced model of the hybrid CNN and LSTM model [5]. It extracts robust features using a concatenation of two word-embedding vectors and a set of CNN layers. Then it learns the order dependencies of extracted features in both directions. Finally, it uses a set of auxiliary output layers to discriminate classes from each other. The proposed model architecture is shown in Fig. 2.

5.1.1 Word Vector Representations

The post padding technique is used to pad articles into 984, which is the maximum length of the articles after the pre-processing phase. The tokenizer method in Keras[2] is used to tokenize articles, encode Arabic terms into indices, and compute the frequency of each token in the training and testing sets. The number of tokens in the vocabulary file is 86,241. GloVe model [30] and pre-trained Arabic FastText[3] are used to generate two 300-dimensional word-vector representations using the pre-processed dataset. First, the GloVe model uses the frequency of each token to generate a GloVe representation. Second, the pre-trained Arabic FastText vectors are mapped to tokens in the proposed vocabulary to generate FastText vectors as shown in Fig. 2.

5.1.2 Inputs and the Embedding Layers

Unlike the model in [5], the length of input sequences is 984 to utilize as much as possible of the information in the news articles that contain long text and to avoid losing details. The proposed model utilizes features from two word-embedding layers that map the GloVe and FastText vectors to encode articles. Both embedded layers are concatenated and fed to the model for feature robustness.

5.1.3 Extracted Features and the Dependency Learning

Three convolution layers of two dimensions are used to extract more robust features [24]. The pooling layer was eliminated to preserve the spatial dimension of features, where the model extracts more features without increasing the computational complexity [34]. The ELU-gate unit is an alternative to the pooling layer. It determines which features are absolutely required and which ones are required to preserve its linearity.

Furthermore, the two branches are assembled using a multiplication operator and fed into the third convolution layer to choose the most relevant features of the ELU and linear features. The number of filters is increased to 256 for better performance [35]. The proposed model uses Bi-LSTM layer instead of LSTM layer to perform feature extraction and learn sequences in both directions.

The number of units in Bi-LSTM is 128, which produces better results [35]. Hence, the number of feature maps is 256 because it is a bi-directional approach. Different activation functions were used to test the Bi-LSTM layer. Consequently, the training process is tuned using ELU and RELU activation functions for forward and recurrent steps, respectively. Similar to the model in [24], the layers of both CNN and Bi-LSTM are followed by l2 regularization, batch normalization, and dropout layers to reduce the effect of overfitting.

5.1.4 Output Layers and Loss Functions

The design of the output layer is inspired by the solutions proposed in [35, 36], where multiple output layers are used. The proposed model is composed of one primary layer and a set of auxiliary layers. The number of auxiliary layers equals the number of classes. Hence, the number of auxiliary layers for the binary classification and multi-class classification tasks are 2 and 3, respectively. The auxiliary outputs are fully connected layers that use Sigmoid activation functions to classify inputs into classes. The first auxiliary output is a binary classification to detect the samples labeled by zero, the second output is to classify samples labeled as 1, and so on. Inspired by the squeeze-and-excitation networks model [5, 37], the binary classification task uses the Sigmoid activation function to enhance and improve the accuracy. The auxiliary branches improve the performance by allowing the model to discriminate each class using binary classification. Each binary branch feeds its output to the corresponding loss function and the primary output layer. The loss function for each branch is the mean squared error using Keras library.[4]

[2] https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer.

[3] https://fasttext.cc/docs/en/pretrained-vectors.html.

[4] https://www.tensorflow.org/api_docs/python/tf/keras/losses/MeanSquaredError.

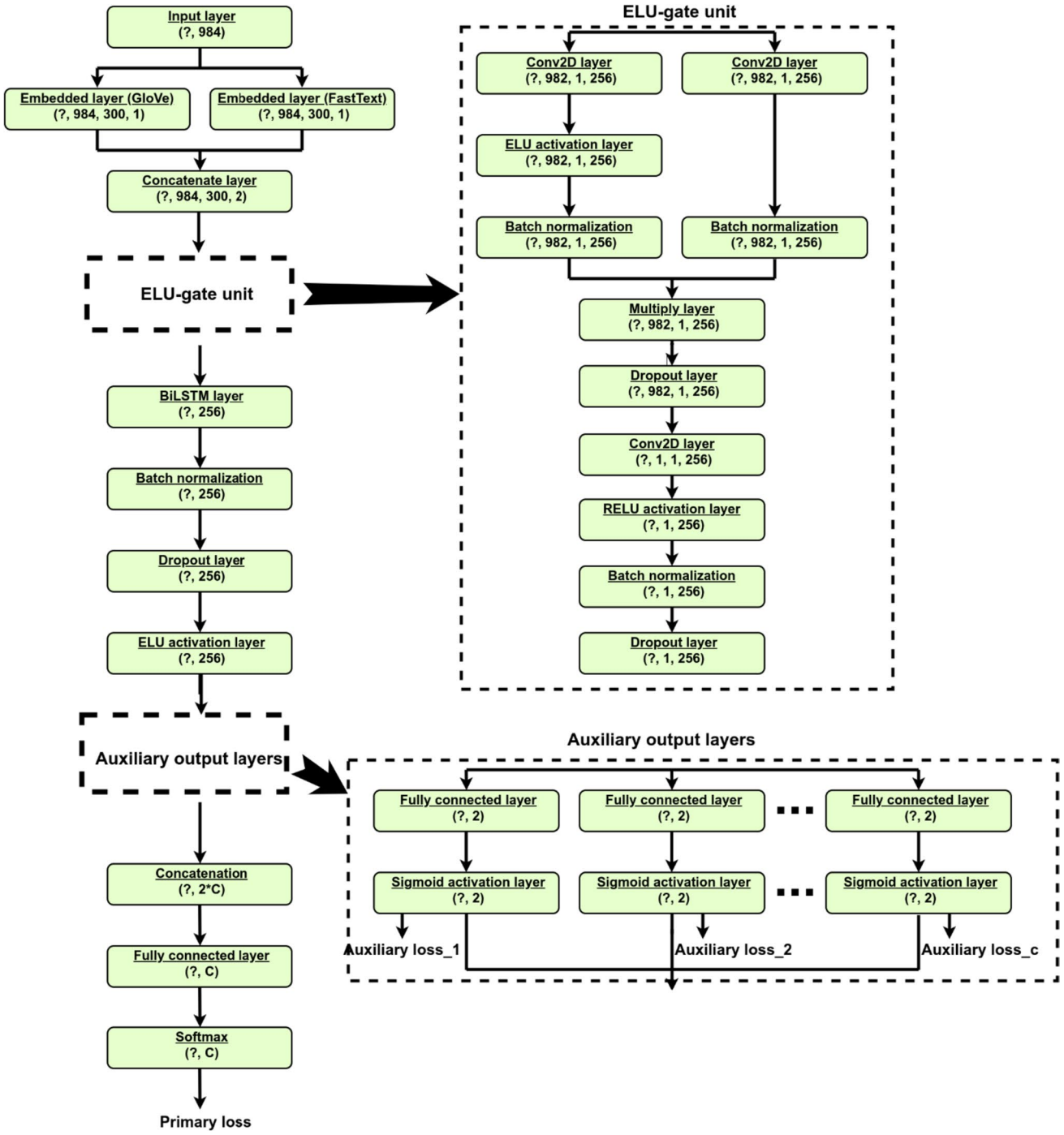

Fig. 2 The proposed architecture

In the multi-class classification task, we have noticed that the auxiliary branches suffer from imbalanced data. Thus, the classes weights are computed for the loss functions using the library Scikit-learn.[5] The weights of classes are computed based on the samples' distribution for both tasks.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ut ils.class weight .compute class weight .html.

Table 1 The accuracy of the multi-class classification using AFND dataset

Model	Valid	Test
Nasir et al. [5]	70.04%	70.31%
Khalil et al. [23]	70.9%	71.0%
Bi-LSTM	38.1%	38.1%
CNN	37.99%	38.1%
Proposed model using GloVe representation	77.15%	77.40%
Proposed model using FastText representation	76.53%	76.64%
Proposed model without auxiliary outputs	77.56%	77.93%
Proposed model	77.59%	77.75%

Table 2 The accuracy of the binary classification using AFND dataset

Model	Valid	Test
Nasir et al. [5]	83.21%	82.80%
Khalil et al. [23]	79.0%	78.3%
Bi-LSTM	53.43%	53.23%
CNN	53.24%	53.33%
Proposed model using GloVe representation	87.29%	87.49%
Proposed model using FastText representation	86.50%	86.76%
Proposed model without auxiliary outputs	87.05%	87.26%
Proposed model	87.23%	87.39%

The output Sigmoid probabilities for each branch are concatenated and fed to the primary output layer which consists of a fully connected layer, categorical cross-entropy,[6] and Softmax activation function.

6 Experiments and Results

6.1 Experiment Settings

The following features of the computer node utilized for the research experiments are listed: an Intel processor with eight cores, 16GB of RAM, and an Nvidia Quadro P4000 graphics card with 12GB of RAM. The training process used the Adam optimizer and the 0-fold cross-validation method [38]. We stopped the training process when the model converged, where the used learning rate was 0.001. The loss weights of the primary output and auxiliary outputs are 1.0 and 0.1, respectively. The number of hyper-parameters in the proposed model is approximately 117 million, which is considered to be very high. However, current computing resources can deal with such numbers in an acceptable performance.

6.2 Experiments and Evaluation

Tables 1 and 2 demonstrate that for both binary and multi-class classification tasks, the proposed model performed well when employing the AFND dataset. As shown in the tables, a slight improvement in the accuracy can be achieved, when using the GloVe representation vector rather than using the concatenation of the GloVe and FastText vectors for the binary classification. Moreover, a good accuracy improvement is achieved when using the concatenation of the two word-representation vectors for the multi-class classification task. The auxiliary output layers have improved the accuracy in both tasks using the valid set. In addition, the auxiliary output layers have improved the performance of the binary classification problem in the test set and slightly reduced it for the multi-class classification problem.

The model is compared to two of the best performing related solutions [5, 23]. Compared with [5], the accuracy of our model is enhanced by 7.66% for the binary classification task and 6.97% for the multi-class classification task. Furthermore, in comparison with the work in [23], for binary and multi-class classification, the accuracy has improved by 8.66% and 6.62%, respectively. Nevertheless, the computation complexity is higher than the model in [5], where the number of hyper-parameters is increased by 109 million, approximately. Nevertheless, the proposed model has achieved better accuracy as a result of extracting more robust features that contribute to the improvement in the accuracy and reducing the effect of the misclassification as shown in Figs. 3 and 4.

Figure 3 represents the confusion matrices of the reference and proposed models in the multi-class classification task. The correct predictions for all classes are increased in the proposed model, which is indicated by a better performance. The high prediction accuracy for the not-credible class is the reason behind the improvement in the accuracy for the binary classification task as shown in Fig. 4, where zero indicates the not-credible class and one indicates the credible class (labels). As shown in the figure, both the proposed and the reference models have similar number of correct predictions for the credible class as reported in [5].

Table 3 and 4 present the performance of the proposed model in terms of the macro precision, macro recall and macro F1-score. The results show that the proposed method is consistent and provides a high confidence for the classification tasks.

In addition, Table 5 and 6 present the accuracy of the different classes (i.e., credible, not-credible, and undecided) for the two classification tasks. Table 5 shows

[6] https://keras.io/api/losses/probabilistic losses/#categorical crossentropy-function.

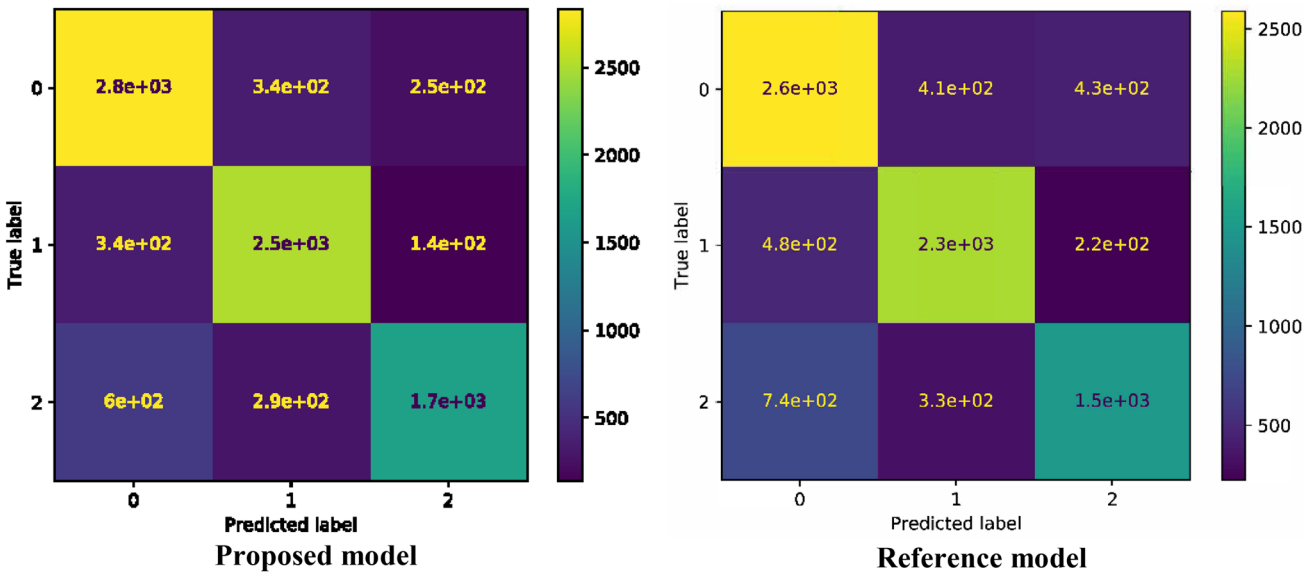

Fig. 3 The confusion matrices for multi-class classification of the proposed and reference models

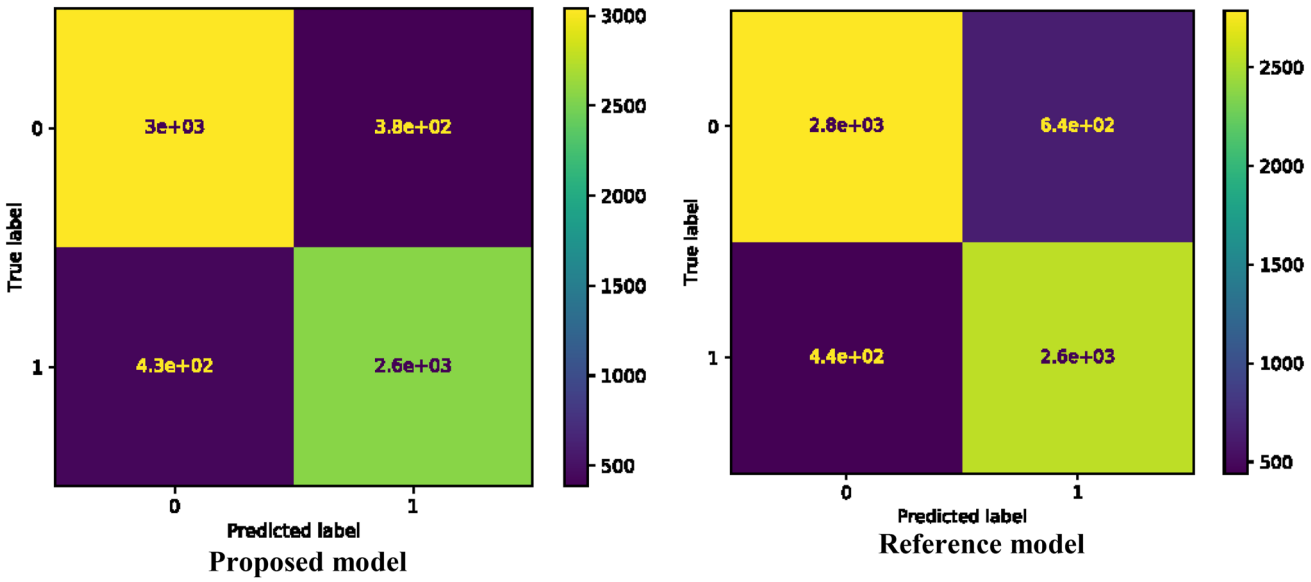

Fig. 4 The confusion matrices for binary classification of the proposed and reference models

Table 3 Performance metrics of the multi-class classification

Model	Macro recall	Macro precision	Macro F1-score
Proposed model using GloVe representation	76.91%	77%	77%
Proposed model using FastText representation	76.17%	77%	76%
Proposed model without auxiliary outputs	77.41%	78%	78%

that the undecided news articles (class 2) have the lowest accuracy which means they are misclassified with higher rate than articles in class 0 and class 1 categories. Credible and not-credible news articles have similar accuracy for the multi-class classification. However, the accuracy of credible articles is higher than not-credible articles in the binary classification.

Table 4 Performance metrics of the binary classification

Model	Macro recall	Macro precision	Macro F1-score
Proposed model using GloVe representation	87.32%	88%	87%
Proposed model using FastText representation	86.58%	87%	87%
Proposed model without auxiliary outputs	87.12%	87%	87%

Table 5 The accuracy of the articles' classes in the multi-class classification

Model	Class 0 (credible)	Class 1 (not-credible)	Class 2 (undecided)
Proposed model using GloVe representation	79.71%	81.68%	69.35%
Proposed model using FastText representation	79.47%	79.5%	69.55%
Proposed model without auxiliary outputs	80.7%	81.8%	69.74%

Table 6 The accuracy of the articles' classes in the binary classification

Model	Class 0 (credible)	Class 1 (not-credible)
Proposed model using GloVe representation	89.8	84.85
Proposed model using FastText representation	89.36	83.79
Proposed model without auxiliary outputs	89.11	85.13

7 Conclusion

In this paper, we have proposed a machine learning model for the detection of fake news articles using Arabic dataset. CNN and Bi-LSTM techniques were used in the proposed model to extract more robust features. The proposed model reduced the misclassification problem and increased the accuracy by more than 7%, on average, for binary classification and multi-class classification. Although the accuracy is increased when using the concatenation of two word-embedding vectors for the multi-class classification task, it is reduced for the binary classification task. Furthermore, the accuracy is increased using auxiliary outputs for both tasks using the valid set. Moreover, the auxiliary outputs improved accuracy using the test set for binary classification. However, the accuracy is reduced when using the test set for the multi-class classification task. Hence, future machine learning methods are needed in order to achieve a higher improvement in accuracy for both multi-class classification and binary classification that focus on Arabic fake news detection.

Acknowledgements We would like to thank Jordan University of Science and Technology for providing the computation and communication resources that enabled us in collecting the dataset and conducting the experiments for this research project.

Authors' Contributions AK: Data curation, Investigation, Formal analysis, Software, Validation, Writing—original draft; MJ: Resources, Conceptualization, Validation, Methodology, Supervision, Writing—review & editing, Funding acquisition, Project administration; MA: Supervision, Conceptualization, Writing—review & editing, Funding acquisition.

Funding This research was supported by the Deanship of Research at Jordan University of Science and Technology (grant number 20210011/832-2020), and in part by Zayed University Research Office, Research Incentives Grant Number R20089.

Availability of Data and Materials The dataset is available at Mendeley under the titled ''Arabic Fake News Dataset (AFND)", URL: https://data.mendeley.com/datasets/67mhx6hhzd/1.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Consent for Publication Not applicable.

Ethics Approval and Consent to Participate Not applicable.

References

1. Alzanin SM, Azmi AM. Rumor detection in Arabic tweets using semi- supervised and unsupervised expectation–maximization. Knowl Based Syst. 2019;185: 104945.
2. Al-Yahya M, Al-Khalifa H, Al-Baity H, AlSaeed D, Essam A. Arabic fake news detection: comparative study of neural networks and transformer-based approaches. Complexity. 2021;2021:1–10.
3. Antoun W, Baly F, Achour R, Hussein A, Hajj H. State of the art models for fake news detection tasks. In: 2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT). 2020. p. 519–524.
4. Jardaneh G, Abdelhaq H, Buzz M, Johnson D. Classifying Arabic tweets based on credibility using content and user features. In: 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE; 2019. p. 596–601.
5. Nasir JA, Khan OS, Varlamis I. Fake news detection: a hybrid cnn-rnn based deep learning approach. Int J Inf Manag Data Insights. 2021;1(1): 100007.
6. Bahad P, Saxena P, Kamal R. Fake news detection using bi-directional lstm-recurrent neural network. Procedia Comput Sci. 2019;165:74–82.
7. Sahoo SR, Gupta BB. Multiple features based approach for automatic fake news detection on social networks using deep learning. Appl Soft Comput. 2021;100: 106983.
8. Faustini PHA, Covoes TF. Fake news detection in multiple platforms and languages. Expert Syst Appl. 2020;158: 113503.
9. Seo Y, Seo D, Jeong C-S. Fander: fake news detection model using media reliability. In: TENCON 2018–2018 IEEE region 10 conference. IEEE; 2018. p. 1834–1838.
10. Alhindi T, Alabdulkarim A, Alshehri A, Abdul-Mageed M, Nakov P. Arastance: A multi-country and multi-domain dataset of arabic stance detection for fact checking. 2021. arXiv preprint arXiv:2104.13559.
11. Ghanem B, Glavas G, Giahanou A, Ponzetto SP, Rosso P, Pardo FMR. UPV-UMA at CheckThat! Lab: verifying Arabic claims using a cross lingual approach. In: CLEF 2019: working notes of CLEF 2019—conference and labs of the evaluation forum, Lugano, Switzerland, vol. 2380. 2019. p. 1–10.
12. Deepak S, Chitturi B. Deep neural approach to fake-news identification. Procedia Comput Sci. 2020;167:2236–43.
13. Bang Y, Ishii E, Cahyawijaya S, Ji Z, Fung P. Model generalization on covid-19 fake news detection. 2021. arXiv preprint arXiv:2101.03841
14. Mahlous AR, Al-Laith A. Fake news detection in Arabic tweets during the covid-19 pandemic. Int J Adv Comput Sci Appl. 2021. https://doi.org/10.14569/IJACSA.2021.0120691.
15. Sabbeh SF, Baatwah SY. Arabic news credibility on twitter: an enhanced model using hybrid features. J Theor Appl Inf Technol. 2018;96(8):2327–38
16. Khouja J. Stance prediction and claim verification: an Arabic perspective. 2020. arXiv preprint arXiv:2005.10410.
17. Khalil A, Jarrah M, Aldwairi M, Jaradat M. Afnd: Arabic fake news dataset for the detection and classification of articles credibility. Data Brief. 2022;42: 108141. https://doi.org/10.1016/j.dib.2022.108141.
18. Muaad AY, Jayappa H, Al-antari MA, Lee S. ArCAR: a novel deep learning computer-aided recognition for character-level Arabic text representation and recognition. Algorithms. 2021;14(7):216.
19. Alhawarat M, Aseeri AO. A superior Arabic text categorization deep model (satcdm). IEEE Access. 2020;8:24653–61.
20. Elnagar A, Al-Debsi R, Einea O. Arabic text classification using deep learning models. Inf Process Manag. 2020;57(1): 102121.
21. Einea O, Elnagar A, Al Debsi R. Sanad: single-label Arabic news articles dataset for automatic text categorization. Data Brief. 2019;25: 104076.
22. Kanan T, Kanaan R, Al-Dabbas O, Kanaan G, Al-Dahoud A, Fox E. Extracting named entities using named entity recognizer for arabic news articles. Int J Adv Stud Comput Sci Eng. 2016;5(11):78–84.
23. Khalil A, Jarrah M, Aldwairi M, Jararweh Y. Detecting Arabic fake news using machine learning. In: 2021 second international conference on intelligent data science technologies and applications (IDSTA). IEEE; 2021. p. 171–177.
24. Kim J, Jang S, Park E, Choi S. Text classification using capsules. Neurocomputing. 2020;376:214–21.
25. Aldwairi M, Alwahedi A. Detecting fake news in social media net- works. Procedia Comput Sci. 2018;141:215–22. https://doi.org/10.1016/j.procs.2018.10.171. (**The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops**).
26. Mukhaini FA, Abdoulie SA, Kharuosi AA, Ahmad AE, Aldwairi M. False: fake news automatic and lightweight solution. In: 2022 IEEE international conference on industry 4.0, artificial intelligence, and communications technology (IAICT). 2022. p. 49–54. https://doi.org/10.1109/IAICT55358.2022.9887471
27. Al Zaatari A, El Ballouli R, ELbassouni S, El-Hajj W, Hajj H, Shaban K, Habash N, Yahya E. Arabic corpora for credibility analysis. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16). 2016. p. 4396–4401.
28. Helwe C, Elbassuoni S, Al Zaatari A, El-Hajj W. Assessing Arabic weblog credibility via deep co-learning. In: Proceedings of the fourth arabic natural language processing workshop. 2019. p. 130–136.
29. Antoun W, Baly F, Hajj H. Arabert: transformer-based model for Arabic language understanding. 2020. arXiv preprint arXiv:2003.00104.
30. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. p. 1532–1543.
31. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. 2021. https://nlp.stanford.edu/projects/glove/. Accessed 1 September 2021.
32. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist. 2017;5:135–46.
33. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. J Big Data. 2021;8(1):1–74.
34. He K, Sun J. Convolutional neural networks at constrained time cost. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 5353–5360.
35. Khalil A, Jarrah M, Al-Ayyoub M, Jararweh Y. Text detection and script identification in natural scene images using deep learning. Comput Electr Eng. 2021;91: 107043.
36. Qiu, Y., Zhang, J., Zhou, J.: Improving gradient-based adversarial training for text classification by contrastive learning and auto-encoder. 2021. arXiv preprint arXiv:2109.06536

37. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7132–7141.
38. Berrar D. Cross-validation. Encycl Bioinform Computat Biol. 2019;1:542–5.