future internet

Article

Attention-Enriched Mini-BERT Fake News Analyzer Using the Arabic Language

Husam M. Alawadh [1], Amerah Alabrah [2], Talha Meraj [3,*] and Hafiz Tayyab Rauf [4]

[1] Department of English Language and Translation, College of Languages and Translation, King Saud University, Riyadh 11451, Saudi Arabia
[2] Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[3] Department of Computer Science, COMSATS University Islamabad—Wah Campus, Wah Cantt 47040, Pakistan
[4] Centre for Smart Systems, AI and Cybersecurity, Staffordshire University, Stoke-on-Trent ST4 2DE, UK
* Correspondence: talhameraj32@gmail.com

Abstract: Internet use resulted in people becoming more reliant on social media. Social media have become the main source of fake news or rumors. They spread uncertainty in each sector of the real world, whether in politics, sports, or celebrities' lives—all are affected by the uncontrolled behavior of social media platforms. Intelligent methods used to control this fake news in various languages have already been much discussed and frequently proposed by researchers. However, Arabic grammar and language are a far more complex and crucial language to learn. Therefore, work on Arabic fake-news-based datasets and related studies is much needed to control the spread of fake news on social media and other Internet media. The current study uses a recently published dataset of Arabic fake news annotated by experts. Further, Arabic-language-based embeddings are given to machine learning (ML) classifiers, and the Arabic-language-based trained minibidirectional encoder representations from transformers (BERT) is used to obtain the sentiments of Arabic grammar and feed a deep learning (DL) classifier. The holdout validation schemes are applied to both ML classifiers and mini-BERT-based deep neural classifiers. The results show a consistent improvement in the performance of mini-BERT-based classifiers, which outperformed ML classifiers, by increasing the training data. A comparison with previous Arabic fake news detection studies is shown where results of the current study show greater improvement.

Keywords: Arabic language; fake news analyzer; mini-BERT; BERT classifier; transformers

check for updates

Citation: Alawadh, H.M.; Alabrah, A.; Meraj, T.; Rauf, H.T. Attention-Enriched Mini-BERT Fake News Analyzer Using the Arabic Language. *Future Internet* **2023**, *15*, 44. https://doi.org/10.3390/fi15020044

Academic Editors: Paulo Quaresma and Wei Yu

Received: 17 November 2022
Revised: 24 December 2022
Accepted: 10 January 2023
Published: 22 January 2023

1. Introduction

The number of Internet users is rising, and social media are becoming the central point of interactivity [1]. The distance between people is reduced via social media platforms such as Twitter, Facebook, and Instagram. The open sharing of content on social media can become dangerous, as in the case of sharing fake news, and creates a number of severe problems such as character elimination [2], fake political favors and negativities [3], and psychological impatience [4]. There is no single definition of fake news, although it often refers to biased information that is spread for gains in politics, business, etc. [5].

Due to the spread of fake news, people's opinions about anything can be changed, which can have serious negative and unwanted consequences. In the 2016 US elections, most US nationals used social media, where deceptive news about political parties changed opinions. Some studies reported that many fake accounts were created to conduct a false campaign against targeted political parties [6]. The news about Pope Francis' endorsement of Donald Trump gained almost one million views. Similarly, YouTube videos containing different misinformed content about COVID-19 gained millions of views. Fake news about any celebrity death or marriage also gains millions of engagements [7].

After 2016, fake news became a trend in many countries. Even the Pew Research Center, which is a nonpartisan social fact checker of US nationals, showed that adults in their own country are involved in the spread of fake news. Of adults on social media, 64% react negatively to the events happening in their country [8]. It seems that fake news is receiving more attention, and this could be due to the propaganda used in this news. Authentication becomes less prominent in societies when people believe fake news. Sometimes, believing or accepting deceptive news is necessary for people to be accepted into a particular network [8]. However, the human ability to detect fake news is not very satisfactory.

According to a psychological experiment, in more than 100 experiments with 1000 participants, only 54% of the accurate identification of fake news was achieved [9]. Therefore, it is necessary to identify fake news automatically, without using the human eye and mind. Some public resources are available that focus on the efficient and wide-scale authentication of social and online news. These include sites such as Snopes and PolitiFact.

Few studies have reported on the existence of fake news in the years before the common era (BCE); the concept was reinitiated when printed media were launched [10]. However, of the paradigm shifts caused by news-spreading media, the arrival of social media sites has most allowed for the rise in fake news. Social media exponentially increase the dissemination of news [11]. Fake news is produced using facts and figures about any personality, object, or event that render its users uncertain about the truth. Sometimes, it leads to mental health problems, and deceptive information about the reputation of some companies can result in significant reduction in business for these companies [12].

Due to the rising challenge of automatically detecting fake news, recent studies have focused on these issues. Social and other news content has recently attracted attention [13]. Arabic news channels and social media usage are also rising in countries facing challenges regarding fake news. The Arabic language is spoken on news channels and social media pages throughout the middle east and north Africa. However, creating an authentic and balanced dataset that is collected in real time with appropriate features is quite a major challenge [14].

Few studies have used private and public datasets. The proposed study also uses a public Arabic news dataset that was published recently on the basis of the real-time data of reliable and unreliable news. The spread of fake news and rumors in the Arab region is a similar magnitude; for example, during the COVID-19 pandemic [15], people believed social media misinformation about COVID-19 symptoms and, later, the vaccines [16]. These were all spread using social media networks such as Twitter and Facebook. Therefore, it is imperative to devise an automated solution that can identify Arabic fake news.

Investigations of fake news concluded that it contains deceptive cues, an informal style of news, particular writing patterns, and false sentiments [17]. Fake news has previously been detected by focusing on sentiments, linguistic cues, and style. However, the increasing use of deep learning (DL) models in sentimental and discourse analysis studies inspired us to create a DL-based solution. The proposed study contributes to this in the following ways:

- The Arabic news analyzer originality check was generated with the help of Arabic semantics and a BERT classifier.
- The original, labeled, and augmented Arabic datasets using DL and ML classifiers were evaluated to strengthen the performance of Arabic fake news analyzers compared to previous studies.
- ML and DL methods were used to compare the performance of the Arabic fake news analyzer, where the DL method assisted in terms of attention masks to pay more attention to the region of interest. This was proven to be a more robust fake news analyzer.

The rest of the article is divided into five sections: Section 2, related work; Section 3, proposed methods; Section 4, results and discussion; Section 5, conclusions with future suggestions. Many of the acronyms used in this study are outlined in Table 1.

Table 1. Different acronyms used in this study.

Acronyms	Description
AUC	Area under curve
BERT	Bidirectional encoder representations from transformers
CNN	Convolutional neural network
DL	Deep learning
GRU	Gradient recurrent unit
LR	Learning rate
LSTM	Long short-term memory
ML	Machine learning
NLP	Natural language processing
SVM	Support vector machine
TF-IDF	Term frequency–inverse document frequency

2. Related Work

Automated solutions with artificial-intelligence-based rules were previously proposed that used ML and DL methods to detect fake news or text. The Arabic language dataset was collected using different web-scraping methods or public datasets. These approaches achieved significant results. Sarcasm was detected in two Arabic news datasets: (1) misogyny and (2) Abu-Farah. Seven different ML classifiers achieved the highest accuracy using the BERT method. Binary and multiclass classifications were performed where 91% accuracy was achieved for the binary class, and 89% for multiclass problems using the same misogyny dataset. Sarcasm detection was also applied using the BERT method, and outperformed by 88% for binary and 77% for multiclass classification [18].

As discussed in the introduction, Arabs also seem to fall victims of fake news regarding emerging issues and COVID-19 was a prime example. An ensemble approach using DL was applied to this topic. Twitter data related to COVID-19 were used to identify fake and real news. The results showed the proposed ensemble method outperformed other approaches [19]. Users of social media or bloggers have the freedom to post on their feeds, allowing for less scrutiny on news and as a result less credibility. Therefore, website ranking and authentication are important ways to establish the credibility of the news posted on these sites.

The ranking of news websites was used to compare the news posted on these sites with authenticated news sites, and a new accuracy score was computed. A new method was used to authenticate the sites using the proposed score criteria. The top1 cosine method was used to obtain a similarity index between different news items. The TF-IDF method was applied to the news, and the half 50% score was calculated using the top1 cosine method. A new term of accuracy was used that was compared with previous studies to prove that it was not considered before [20].

Rumors on Arabic tweets were detected by a study using a gradient boosting approach. Rumors are also disseminated extensively on social media sites and are also considered to be fake news. Twitter data on rumors and nonrumors were used. Content-, user-, and topic-based features were extracted after the preprocessing of Twitter data. The finalized data were then applied to classical ML methods. On 60% of the training data, the eXtreme Gradient Boost (XG-Boost) had 97.18% accuracy in classifying the data from rumors [21]. A summary of these studies is shown in Table 2.

Table 2. Summary of fake news detection methods applied in 2022.

References	Dataset	Methods	Results
[18]	Misogyny, sarcasm	BERT and other classical ML methods	Misogyny: Binary = 91%, Multi = 89% Sarcasm: Binary = 88%, Multi = 77%
[19]	Twitter data of COVID-19 tag news	Ensemble DL model	Weighted F1 score = 0.99
[20]	Different datasets	Website ranking + Tf-IDF scores used to calculate the proposed news' accuracy	-
[21]	Arabic Tweeter data of rumors and nonrumors	Content-, user-, and topic-based features with machine learning classifiers	Accuracy using XG-boost = 97.18%
[22]	ISOT	Statistical, contextual features with machine learning classifiers, BERT, GRU, and LSTM	Highest accuracies: GRU = 0.988, LSTM = 0.991
[23]	Kaggle dataset	Soft voting classifier, SVM, LR, Naïve Bayesian, and FridSearchCv optimization	Highest accuracy achieved by ensemble method = 93% accuracy

Real and fake news detection was applied to big data that contained political, governmental, US, Middle East, and general news. After preprocessing, the dataset was cleaned, and contextual and statistical features were extracted and individually embedded into four different ML classifiers: SVM, random forest, naïve Bayesian, and decision tree. The BERT model, LSTM, and GRU methods were also applied, and the proposed stacked models of LSTM and GRU achieved the highest accuracy compared to previous studies. Accuracies of 0.991 and 0.988 were achieved by the GRU and LSTM models, respectively [22].

The soft voting classifier was used by a few studies, such as [23]. The authors used it to aggregate classical ML classifiers, such as SVM, LR, and naïve Bayesian. During the training of these algorithms, optimization method FridSearchCV was used. Kaggle provided a dataset of fake and real news. The results showed that the ensemble approach achieved the best results, with 93% accuracy, and an F1 score with 94% precision and 92% recall values.

In all the above discussions about recent studies on Arabic and English fake news detection methods, Arabic news was mostly collected with individual scraping methods or manual collection from news sites. However, these methods have limitations when finding and using a balanced, real, and authenticated dataset with associated ground-truth annotations. Furthermore, very few studies were proposed on Arabic fake news detection as compared to English fake news detection. Therefore, a balanced, authenticated, and robust approach is needed to enhance automated Arabic fake news detection methods or solutions.

3. Materials and Methods

Many NLP-based sentiment analysis approaches have been proposed as a solution to different real-life case studies that are revolutionizing the world by using different improved artificial-intelligence-based techniques. In NLP history, sequence-to-sequence models were first used to convert a language text into another language text, and then recurrent neural network (RNN)-based models were used that recurred until the weights had been updated. After that, memory-containing model LSTM was used, which keeps the data recurring, and updates the weights in a particular way. The attention-mask-based mechanism was introduced in 2015 that creates the context vector of a pass-on instance using a weighted sum of the hidden layers of the network. From these mind-storming models and workings, Google introduced attention-mechanism-based transformer models. The main idea of these models is to ensure that attention-enriched input and output retain their dependencies with the recurrence of neural networks. The self-attention of data is used while the embedding is passed to the first encoder block; in the second block, a feed-forward pass to the coming data is completed, and self-attention is used again.

At the end of the network, when the decoder part begins, self-attention to data is carried out again by keeping the original context as it is. Intra-attention is achieved by remaining context-aware to discern the appropriate sentiment of the text. These transforming model variants have been proposed by many studies. However, the Arabic-text-analysis-based BERT was proposed recently and has not been extensively tested on Arabic language datasets.

In this study, the Arabic-language-based transformer model was used to detect fake news on Arabic news channels. At the start of the study, normal preprocessing was used to remove useless words, and then mini-BERT model-acquired data-based preprocessing was carried out. After obtaining unique input tokens and attention masks, training was carried out using the pretrained mini-BERT method. The three data splits were performed on the data, and tested on a mini-BERT classification model. Further, to prove the robustness, the same splits-based classification was performed on classical ML classifiers. All steps are shown in Figure 1. A detailed discussion of each classifier's performance is given in Section 4.

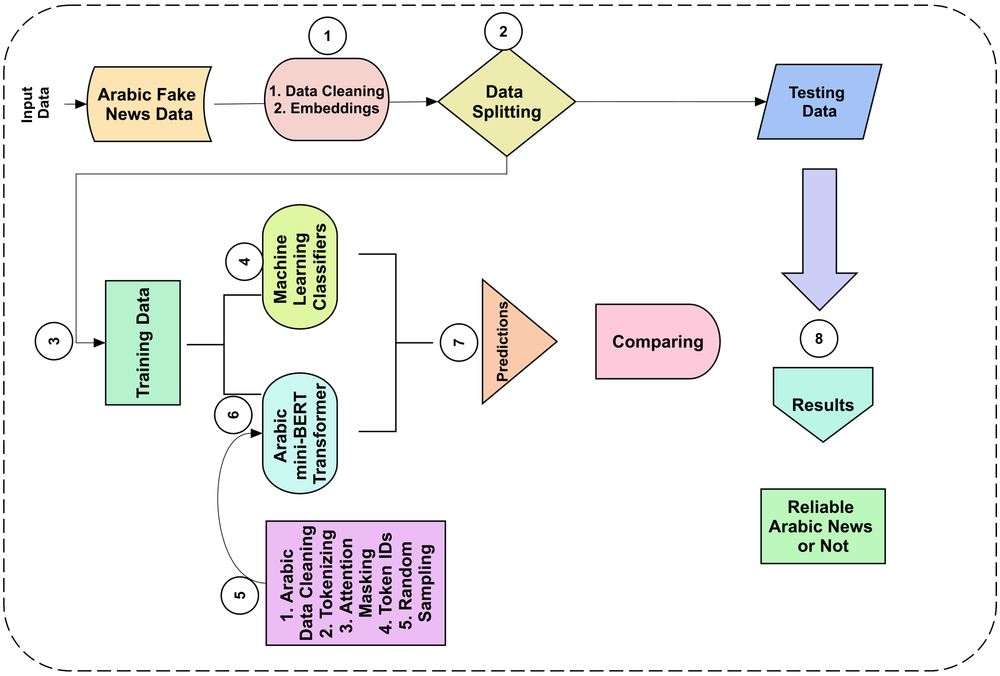


Figure 1. Primary steps of a study conducted on Arabic fake news detection using ML and mini-BERT classifiers.

3.1. Preprocessing Text

Unicode data were processed first, where the Arabic text was freed with URLs, the white spacing, the '@' character, and other useless signs, and then these data were processed using mandatory BERT preprocessing methods. In these steps, the tokenizer from mini-BERT [24] was loaded into the workspace. The reason for using this version of BERT is that it is an Arabic-sentence-based trained model.

The model in the proposed study was trained on 11M parameters with 256 hidden layers, 4 attention heads, and 4 hidden layers. This Arabic version was mainly trained on Google tensor processing units (TPUs) that were designed to train large data tensors. However, the Arabic language is a complex linguistic system; with a slight change in pronunciation, the sentiment of the word changes. Therefore, simple frequency or other features may not fulfil the sentiments in Arabic text. After BERT preprocessing, encoding was applied.

3.2. BERT Preprocessing

In this encoding, special tokens were added to each assigned token of the BERT tokenizer, and the tokens were added at the start and end of a sentence. The maximal length was estimated, and padding was added to shorter documents. Most importantly, attention masks were added against each created token. These contained special weights that were assigned to each token produced by the BERT tokenizer. To obtain a better understanding of the attention masks, they highlighted the region of interest or main sentiment. This shows the sentiment given in the text. Attention masks add a feature vector that corresponds to the input keywords and embedding indices. The attention mask vector contains 0 and 1 values, in which 1 indicates the need to select a corresponding index in the embedding vector to pay attention, and 0 indicates that the corresponding keyword for attention should not be selected. In this way, many of the useless keywords in embeddings were not used while training the BERT model. This not only renders the model less time-consuming, but also means that more attention is paid to more prominent keywords in documents. However, the original work was proposed by Google and is given on the Hugging Face website. The pretrained BERT model needs to add a few layers to perform different tasks of text classification, sentiment analysis, and many other tasks.

These are bidirectional transformers models that were originally produced for use on unlabeled data. They work on attention masks to keep them aware of the context. To summarize how this method performs, random samples of loaded data and their masking are established, predictions are given for those words using their context, and the ID of the predicted word is output. To ensure awareness of the forward and backward contexts, both ways of obtaining the context of a word are used. This joins with the next word in the sentence, and the whole context of the given instance is thus covered. It is also important that this is only available in Pytorch. A better and ongoing way is to improve DL in computer vision and NLP tasks.

3.3. ML Classification

To prove the robustness of the mini-BERT transformer model, various data splits and ML classifier-based comparisons were performed. Classical models, such as decision tree, naïve Bayesian, support vector machine classifier, and random forests [25], were applied. A decision tree was applied with parameters such as the 'gini' method used to validate the quality of splits, and a maximal feature (n) was given where other parameters were set as default. Naive Bayesian was given for the multimodel where parameters were set as alpha = 1, and class priority was set as None. The support vector machine was given with the probability set as true, kernel set to linear, gamma = 0.001, and C = 20. Random forest classifier parameters were set at 100 estimators; the 'gini' method was given as a criterion. The holdout validation methods for splitting training and testing datasets were applied and fed to these classifiers to validate their performance.

Improved results were shown compared to those of the transfer-learning-based model of mini-BERT for Arabic fake news detection. The reason to use these classifiers was their popularity and their excellent performance in previous studies on NLP tasks. These classifiers' predictions behave similarly to each other and are satisfactory in Arabic text.

Mini-BERT Classification

Preprocessed attention masks and tokenized IDs were given to the Pytorch tensor data. The backend environment was set to Pytorch while performing the experiment. This provided us with a random sampler for validation data, the training data loader, and the torch tensor for encoding the specific labels. For training data loading, random samplers, attention masks, and corresponding labels, the tensors provided by the Pytorch utility libraries are all appropriate. The fine-tuned parameters used in the transfer-learning-based classification of Arabic news are presented in Table 3.

Table 3. Fine-tuned parameters of Arabic fake news detection mini-BERT model.

Parameters	Value
Batch size	20
Adam optimizer learning rate	0.00005
Adam optimizer Epsilon	1×10^{-8}
Hidden size	256
Maximal length	280
Epochs	100

After obtaining the training data, the random sample of training data, and their labels, the total batch was loaded to the GPU, and then the gradients and losses were computed and made zero if they contained any garbage values. The forward pass was performed using tensors of attention masks, the input training data, and their corresponding labels. During the forward pass of neural networks, the updated weights are propagated back. To obtain the updated weights, we performed backward passes, and the weights were updated.

For each batch of 20, the batch loss was monitored, and updated weights calculated using validation data were monitored after each epoch to fine-tune the model. The loss was calculated using the ratio of batch loss and the length of the training data. Many BERT versions are available, and some are discussed to provide a comparison with the mini-BERT used by this study.

BERT-large and BERT-base work in similar ways but the large, as in the name, is more frequently used for large datasets and more time is needed to train the model. The other variant, RoBERTa, makes slight changes to the original BERT base by making the batch-size smaller, and the forward context or sentence prediction is removed to reduce the sequence size. Further slight changes are made in the pattern of BERT's original masking phenomenon.

This was frequently tested on different languages and has been more promising results than those of the original BERT, even on unbalanced datasets [26]. One of the most important variants is one that not only reduces the computational power of the original BERT base, but also enhances the results of a few conducted studies. This is known as DistilBERT, was proposed in 2019, and provides a 40% reduction in model size. It is used on many different tasks, such as linguistic knowledge [27], important-words selection [28], and voice shipping assistant [29] tasks.

Similarly, Albert, Albert-base-v2, Electra-small, and BART-large were proposed. These all reduce the original model architecture size and its parameters, change the masking pattern, and reduce the computational and memory usage to enhance the results and render the model optimal when using different languages. However, the Arabic language is extensively richer in grammatical context, and predicting its sentiments is quite complex.

A recently proposed Arabic language model reduction, the convolutional neural network (CNN), is extensively used in computer vision tasks and text recognition tasks using text kernel windows. The conducted study of mini-BERT was combined with CNN and the original BERT model and tested on three different languages (Arabic, Turkish, Greek), outperforming in terms of F1 scores. It mainly aims to obtain the tokenization and classification models used by the conducted studies to detect the reliability of Arabic news in its original language, not in English.

4. Results and Discussion

The three splits in the data were performed and given to ML classifiers and Arabic transfer-learning-based deep neural networks. Four evaluation measures were used in this study to validate the results. These are described in the following section.

4.1. Evaluation Measures

The performance was evaluated on the applied methods using Equations (1)–(4). The true and false positive and true and false negative entities were used in these metrics.

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) * 100 \tag{1}$$

$$Precision = \left(\frac{TP}{TP + FP}\right) * 100 \tag{2}$$

$$Recall = \left(\frac{TP}{TP + FN}\right) * 100 \tag{3}$$

$$F1score = 2 * \left(\frac{recall * precision}{recall + precision}\right) * 100 \tag{4}$$

The accuracy of Equation (1) shows more true positives and negatives than true and false positive and negative instances, where multiplication with 100 shows the percentage of the result. This is a general measure of performance. The precision measure is taken as a true positive over the sum of true and false positives with the multiplication of the whole ratio to 100. This shows the true positive rate of overall performance.

Equation (3) also shows the same true positive predictivity performance regarding instances of false negatives. The final appropriate measure that was not usually applied in previous studies is discussed in Section 2 and Table 1. The F1 score is taken as the precision and recall product over the summation ratio, multiplied by 2 as a constant, and multiplied by 100 to obtain the percentage sign. This shows that the true positive rate was higher than the false positive and negative rates combined. Further, this covers the class imbalance issue if any arises in a given testing dataset.

4.2. Dataset Description

The dataset used in this study [30,31] was split into 90/10, 80/20, and 70/30 parts. In these splits, the first arguments, 90, 80, and 70, show the training data ratio from the total data, whereas the second figures, 10, 20, and 30, show the testing data ratio. The original data, the augmented, data-based number of each class, and the frequency of each class are shown in Table 4.

Table 4. Class frequency of original and augmented data of Arabic fake news.

Dataset	Reliable	Unreliable	Total
Original	100	222	322
Augmented	200	444	644

Table 3 shows the class frequency of each class along with its total. In the original public data, there were 323 instances in total: one instance was removed because it was null but was still labeled. However, the Arabic data in a comma-separated sheet showed nothing but garbage. Therefore, it was saved in the worksheet and loaded into the workspace for digital acquisition.

In text analysis tasks, thousands or millions of instances of data are used, whereas in these data, only 322 instances were given. To remove the annotated data limitation and retain the original annotated labeled data as big data, a random sampling of the same 322 instances was doubled, and any biases were removed. In the original dataset, the ID, categorical and double types of labels are given with their text and title. In the conducted study, the text and its label were used, while others were discarded.

4.3. ML Classification

We used four popular ML classifiers with three different splits using the holdout validation method. The most famous and appropriate measures of accuracy, precision, recall, and F1 score were calculated, and their results remained satisfactory. The summarized results of each split and the four measures are given in Table 5.

Table 5. Classification results of ML classifiers on embedded Arabic fake news detection.

Splits	Methods	Accuracy	Precision	Recall	F1 Score
70/30	Decision tree	84.97	86.76	91.47	89.05
	Random forest	83.93	82.66	96.12	88.88
	Naïve Bayesian	73.57	72.15	98.44	83.27
	Linear support vector	83.93	86.56	89.92	88.21
80/20	Decision tree	93.02	97.59	92.04	94.73
	Random forest	96.12	97.70	96.59	97.14
	Naïve Bayesian	86.04	83.01	100	90.72
	Linear support vector	94.57	97.64	94.31	95.95
90/10	Decision tree	98.43	100	97.5	98.73
	Random forest	98.43	100	97.5	98.73
	Naïve Bayesian	79.68	76.47	97.5	85.71
	Linear support vector	96.87	100	95.0	97.43

In Table 3, we can see that four different classifiers performed differently on three different splits. If we look at the 70/30 split of the given data and the results of the four classifiers, we can see that 84.97% accuracy was achieved for the testing data by the decision tree, while the random forest classifier achieved 83.93%. If we look at the change in the precision values of the two classifiers, they were 86.76 and 82.66%, respectively. These precision values show the precision of the ratio between true positives and the sum of true and false positives. Therefore, the positive instances that were reliable news sources showed quite good results, and, if we compare all four classifiers' precision values, then the decision tree performed the best in a 70/30 split.

The general measure of all these classifiers was also obtained. The decision tree showed the highest accuracy of 84.97%. The recall measure is the ratio between true positives over the summation of true positives and false negatives. If we look at naïve Bayesian results, its accuracy was not as high compared to the other three classifiers, but the recall value is the highest among all four classifiers, with a value of 98.44%.

However, accuracy is a general measure that is not mainly targeted toward class imbalance issues. In the conducted study, the dataset containing class imbalance showed more than double the instances of the negative class. Therefore, the F1 score is a more appropriate measure than the other three measures. In the other three measures, the scores are more varied, whereas if we investigate the F1 score, the scores are nearer to each other. This means that it is a more realistic or appropriate measure of evaluation.

In the second data split with an 80/20 ratio, more data were used in training, at 80%, and fewer data ertr used in testing, at 20%.More training leads to more accurate results. However, the results were better for all evaluation measures. The 93, 96, 86, and 94% results were obtained by the decision tree, random forest, naïve Bayesian, and linear support vector classifiers, respectively. The highest score was achieved by random forest, which was previously the same as the decision tree; this time, however, the random forest outperformed and was a better classifier than the other four methods.

To increase the validity of the results, let us look at the other three measures. The precision values of 97.59 and 97.70% for the decision tree and random forest, respectively, showed quite similar behavior for true positive over true and false positive ratio measures, whereas 83.0 and 97.64% values of precision were obtained for naïve Bayesian and linear support vector machines. We can see that the performance of the naïve Bayesian again dropped compared to the other three classifiers, and the performance of the linear support vector machine classifiers was good, and nearer to the prediction performance of the decision tree and random forest.

As we investigated the previous split of 90/10, the naïve Bayesian recall performance was good as compared to that of other classifiers; the recall value of naïve Bayesian was 100%, the maximum. The final appropriate measure (F1 score) of evaluation also showed the highest score of 97.14% for the random forest classifier, which also achieved the highest accuracy score among all classifiers. This again shows that random forest is the most appropriate classifier for this split of data.

At the 90/10 ratio of data, the models were trained on 90% data and then tested on 10%. In this way, the models were trained on the maximal data and tested on fewer. However, if the evaluation measures values remained similar or justified in each split, this could prove that, if enough data are available for training and testing, then the performance of these classifiers could be satisfactory and valid. Now, let us look at the performance of these classifiers.

The decision tree and random forest again showed accuracy of 98.43%, which was again the highest accuracy value achieved by any of the four classifiers. The precision value outperformed that of the other three classifiers with 100%, except for the naïve Bayesian classifier. However, this time, the recall value remained similar to those of the random forest and decision tree scores of naïve Bayesian, while the linear support vector method obtained a lower value than these classifiers. The last and distinguishing measure (F1 score) showed the same score for random forest and decision tree: 98.73%. This time, the accuracy and F1 scores remained the same, which makes the results more certain.

The area under the curve (AUC) is another measure that is plotted to check how model predictions for different classes of testing data. This uses the probability values given by any classifier to find a cutoff threshold between true positives and false positives. The AUC values for the three splits of data fed to four ML classifiers are shown in Figure 2.

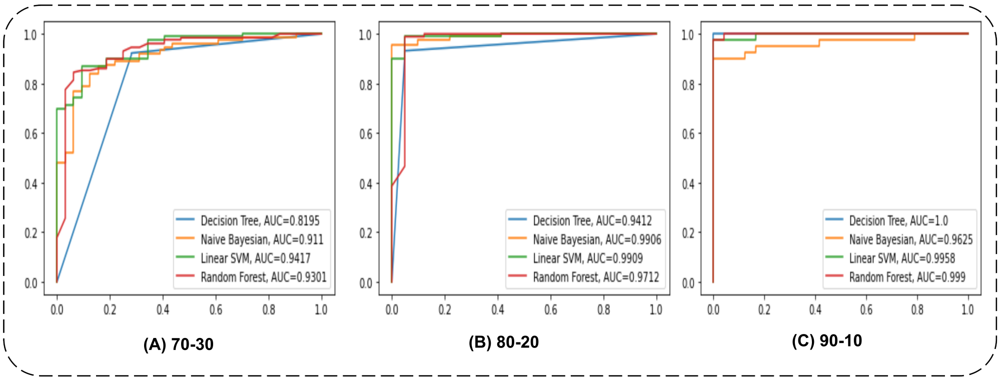
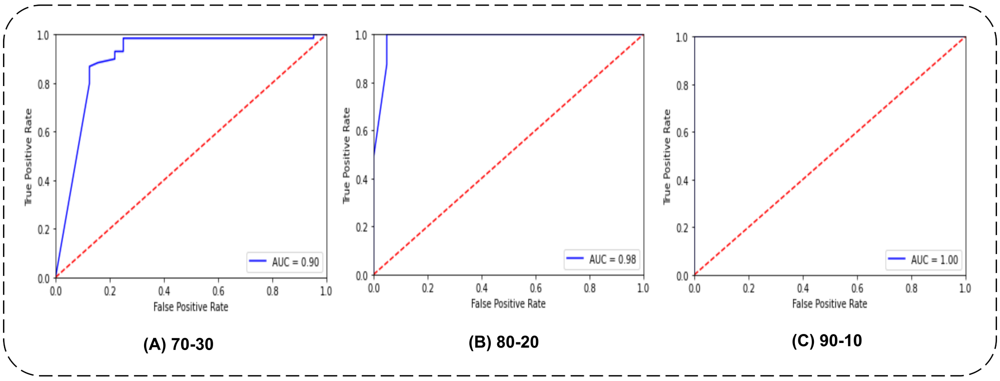


Figure 2. ML classifiers—area under the curve (AUC) for all three splits applied in the proposed study.

4.4. Mini-BERT Transfer Learning Classifier Prediction Results

The pretrained mini-BERT classifier showed the simple deep-learning classifier results obtained with forward and backward propagation. The text CNN and basic-BERT-based Arabic mini-BERT tokenizer and classification model were used in this training. The classifier prediction results for the testing data on various splits are shown in Table 6. The mini-BERT testing data regarding losses and AUC were also monitored and are shown in a large frame in Figure 3.



(A) 70-30

(B) 80-20

(C) 90-10

Figure 3. The testing data predictions of three trained classifiers of mini-BERT-based transfer learning.

With 70% training data, (A) shows an AUC value of 0.90, which is good, although the value of AUC rapidly changes when we train the model on 80% data. The blue curve represents the AUC curve. The last one (C) showed extraordinary results when we used 90% data for training and remained at 1.00 without showing a single point less than 1. This illustrates that when, training data are used less, the validation accuracy gradually increases, while when we increase the training data of mini-BERT classifiers, a simultaneous rise is shown.

Table 6. Arabic Fake News Classification results of Mini-Bert-based Transfer Learning Classifiers using various data splits

Splits	Accuracy	Precision	Recall	F1 Score
70/30	87.04	88.23	93.02	90.56
80/20	96.12	97.70	96.59	97.14
90/10	98.43	100	97.5	98.73

The evaluation measures in the table show that the performance of each split consistently increased when we increased the training of the mini-BERT classifiers. The accuracy measure increased from 87.04 to 98.43%. The precision value increased from 88.23% to 100%, which was also the maximum for the true-positive class predictions. The recall value also changed in ML classification, increasing from 93.02 to 97.5%. The most appropriate measure F1 score showed a 98.73% value, increasing from 90.56%.

After looking at the results of the two types of classifiers, we can see that there was not a large difference between the 90/10 split of data scores. However, between each split, the classifier’s performance kept the F1 score at less than 90%, although in the case of the mini-BERT classifier, the F1 score remained above 90%. Except for cases with data training and more testing, mini-BERT showed more robustness in its results. Therefore, the Arabic sentiment-based trained classifier could be used for other Arabic language tasks.

4.5. Comparison

The conducted study used reliable or unreliable Arabic news and the mini-BERT model. However, the data were recently published and there are not many studies that can be used for comparison. However, the Arabic mini-BERT and classical ML methods results were compared to obtain technically feasible comparisons of similar Arabic data or text.

Table 7. Comparison of fake news data using various machine learning and DL techniques.

References	Purpose	Methods	Results
[21]	Arabic Tweeter data of rumors and nonrumors	Content, user, and topic-based features with ML classifiers	Accuracy using XG-boost = 97.18%
[18]	Misogyny, sarcasm-based classification	BERT and other classical ML methods	Misogyny: binary = 91%, Multi = 89% Sarcasm: Binary = 88%, Multi = 77%
Proposed study	Arabic-language-based fake news detection	Mini-BERT	accuracy = 98.43%, F1 score = 98.73
		ML Classifiers	Best-performing, random forest and decision tree Accuracies = 98.43% F1 score = 98.73%

If we look at the comparison Table 7, we can see recent studies on the Arabic language containing rumor and nonrumor text that was used by content and user-based features, and given to ML classifiers. The XG-boost method classification performed the best, with 97.18% accuracy. Similarly, in another study, BERT and ML classifiers were used on binary and multiclass classification problems using two datasets, and showed satisfactory results.

However, in our comparison with previous solutions, accuracy scores were not emphasized. This is because although the topic may be the same, the dataset is not the same, and, thus, comparison using this measure is irrelevant. However, if we look at a given evaluation measure, the F1 score is a proven more appropriate measure than the other evaluation metrics. Therefore, in terms of the proper use of DL or attention-enriched Arabic fake news detection methods or in terms of the use of appropriate evaluation measures, the conducted study had more satisfactory, robust, and confident results in terms of Arabic fake news detection.

5. Conclusions

Fake news detection using intelligent ML methods is frequently proposed in the English language. However, little scholarly work has been done to address this issue in the Arabic language. Further, creating a large and annotated dataset collection for Arabic news is also a major challenge. However, an annotated dataset of the Arabic language, labeled by experts, was recently published and was used in this study. This study preprocessed this dataset, and applied tokenization and embedding using standard text 2 numeric encodings, which were later fed to ML classifiers. A deep learning approach was also applied in which the Arabic dataset was fed to a mini-BERT transformer model trained in the Arabic language. This provided tokenization, attention masks, and IDs for the Arabic text. The holdout validation scheme was adopted with 70/30, 80/20, and 90/10 splits of data on both ML and DL methods. The performance of ML classifiers deviated between these splits. However, the behavior of the mini-BERT classifiers showed consistency while the training data increased; each type of evaluation measure increased, unlike ML classifiers, which showed more varied behavior on different splits. The highest accuracy was up to 98.43%, and the mini-BERT approach was more valid considering its consistent behavior in performance throughout all splits. The approaches applied in this study showed comparatively better performance. However, the study has limitations regarding the large dataset; if larger datasets are fed to the mini-BERT transformer model, then the performance could change. This could be optimized by fine-tuning the training of BERT models.

In the future, it is recommended that more mini-BERT-based, data-based, Arabic-language sentiment tasks be performed. Large datasets on the Arabic language that are properly and carefully annotated by experts need to be published.

Author Contributions: Conceptualization, H.M.A.; Problme statement: H.M.A.; Methodology, T.M.; software, A.A.; Data Validation, H.M.A.; formal analysis, A.A.; investigation, T.M.; resources, A.A.; data curation, T.M.; writing—original draft, H.T.R.; Editing and Proofreading, H.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: No external funding was received for this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset is publicly available.

Acknowledgments: This research was supported by the Researchers Supporting Project (RSPD2023R799), King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Harrag, F.; Djahli, M.K. Arabic Fake News Detection: A Fact Checking Based Deep Learning Approach. *Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 1–34. [CrossRef]
2. Pulido, C.M.; Ruiz-Eugenio, L.; Redondo-Sama, G.; Villarejo-Carballido, B. A new application of social impact in social media for overcoming fake news in health. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2430. [CrossRef]
3. Maldonado, M.A. Understanding fake news: Technology, affects, and the politics of the untruth. *Hist. Comun. Soc.* **2019**, *24*, 533. [CrossRef]
4. Ozbay, F.A.; Alatas, B. Fake news detection within online social media using supervised artificial intelligence algorithms. *Phys. A Stat. Mech. Its Appl.* **2020**, *540*, 123174. [CrossRef]
5. Meel, P.; Vishwakarma, D.K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Syst. Appl.* **2020**, *153*, 112986. [CrossRef]
6. Lewandowsky, S.; Ecker, U.K.; Cook, J. Beyond misinformation: Understanding and coping with the "post-truth" era. *J. Appl. Res. Mem. Cogn.* **2017**, *6*, 353–369. [CrossRef]
7. Davoudi, M.; Moosavi, M.R.; Sadreddini, M.H. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Syst. Appl.* **2022**, *198*, 116635. [CrossRef]
8. Auxier, B. *64% of Americans Say Social Media Have a Mostly Negative Effect on the Way Things Are Going in the U.S. Today*; Pew Research Center: Washington, DC, USA, 2020.
9. Rubin, V.L. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proc. Am. Soc. Inf. Sci. Technol.* **2010**, *47*, 1–10. [CrossRef]
10. Soll, J.; White, J.B.; Sitrin, S.S.; Carly.; Gerstein, B.M. The Long and Brutal History of Fake News. *Politico Magazine*, 18 December 2016.
11. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
12. Schonfeld, E. Citizen "Journalist" Hits Apple Stock with False (Steve Jobs) Heart Attack Rumor. 2008. https://techcrunch.com/2008/10/03/citizen-journalist-hits-apple-stock-with-false-steve-jobs-heart-attack-rumor (accessed on 15 May 2022).
13. Zhou, X.; Zafarani, R. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explor. Newsl.* **2019**, *21*, 48–60. [CrossRef]
14. Nassif, A.B.; Elnagar, A.; Elgendy, O.; Afadar, Y. Arabic fake news detection based on deep contextualized embedding models. *Neural Comput. Appl.* **2022**, *34*, 16019–16032. [CrossRef]
15. Alotaibi, F.L.; Alhammad, M.M. Using a Rule-based Model to Detect Arabic Fake News Propagation during COVID-19. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 112–119. [CrossRef]
16. Alabrah, A.; Alawadh, H.M.; Okon, O.D.; Meraj, T.; Rauf, H.T. Gulf countries' citizens' acceptance of COVID-19 vaccines—A machine learning approach. *Mathematics* **2022**, *10*, 467. [CrossRef]
17. Zhang, X.; Ghorbani, A.A. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.* **2020**, *57*, 102025. [CrossRef]
18. Muaad, A.Y.; Jayappa Davanagere, H.; Benifa, J.; Alabrah, A.; Naji Saif, M.A.; Pushpa, D.; Al-Antari, M.A.; Alfakih, T.M. Artificial intelligence-based approach for misogyny and sarcasm detection from Arabic texts. *Comput. Intell. Neurosci.* **2022**, *2022*. [CrossRef]
19. Kumar, A.; Singh, J.P.; Singh, A.K. COVID-19 Fake News Detection Using Ensemble-Based Deep Learning Model. *IT Prof.* **2022**, *24*, 32–37. [CrossRef]
20. Mughaid, A.; Al-Zu'bi, S.; Al Arjan, A.; Al-Amrat, R.; Alajmi, R.; Zitar, R.A.; Abualigah, L. An intelligent cybersecurity system for detecting fake news in social media websites. *Soft Comput.* **2022**, *26*, 5577–5591. [CrossRef]
21. Gumaei, A.; Al-Rakhami, M.S.; Hassan, M.M.; De Albuquerque, V.H.C.; Camacho, D. An effective approach for rumor detection of Arabic tweets using extreme gradient boosting method. *Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 1–16. [CrossRef]

22. Amer, E.; Kwak, K.S.; El-Sappagh, S. Context-Based Fake News Detection Model Relying on Deep Learning Models. *Electronics* **2022**, *11*, 1255. [CrossRef]
23. Lasotte, Y.; Garba, E.; Malgwi, Y.; Buhari, M. An Ensemble Machine Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier. *Eur. J. Electr. Eng. Comput. Sci.* **2022**, *6*, 1–7. [CrossRef]
24. Safaya, A.; Abdullatif, M.; Yuret, D. KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona, Spain, 12–13 December 2020; pp. 2054–2059.
25. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
26. Casola, S.; Lavelli, A. FBK@ SMM4H2020: RoBERTa for detecting medications on Twitter. In Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, Barcelona, Spain, 12–13 December 2020; pp. 101–103.
27. Staliūnaitė, I.; Iacobacci, I. Compositional and lexical semantics in RoBERTa, BERT and DistilBERT: a case study on CoQA. *arXiv* **2020**, arXiv:2009.08257.
28. Abadeer, M. Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, Online, 19 November 2020; pp. 158–167.
29. Mozafari, J.; Fatemi, A.; Moradi, P. A method for answer selection using DistilBERT and important words. In Proceedings of the 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, 22–23 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 72–76.
30. Assaf, R. Arabic Fake News Dataset. Available online: https://github.com/RashaAssaf/fake_news_Dtaset (accessed on 15 May 2022).
31. Assaf, R.; Saheb, M. Dataset for Arabic Fake News. In Proceedings of the 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 13–15 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.