ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 11, No. 6, December 2023, pp.56-72

Arabic fake news detection for Covid-19 using deep learning and machine learning

Raad Sadi Aziz [1], Ahmed T. Sadiq [2], Monji Kherallah [1], Ali Douik [3]
[1]University of Sfax, Tunisia
[2]University of Technology, Iraq
[3]University of Sousse, Tunisia

ABSTRACT

When newspapers were the dominant form of conventional media, fake news was widespread. Due to the vast influence of such false news and the growing user reach of technical media sources (TV, Internet, social media, blogs). Humans have become more dependent on the news as they make daily decisions for ensuring the safety of their loved ones and themselves in the wake of COVID-19 becoming a pandemic which has impacted humans all over the world. Fake news, on the other hand, is on the verge of becoming a "second pandemic" or "infodemic," endangering the health of individuals all over the world. Previous research hasn't used fake news detection to coronavirus in Arabic due to the fact that fake news connected to coronavirus is such a recent occurrence. A total of 4 versions of the datasets used in this study have been produced (D0, D1, D2, and D3). To understand the effects of deep learning (DL) and machine learning (ML) techniques on any dataset, a total of 4 datasets were created. Also, the research analyzes them with regard to ML and DL to determine the efficacy of preprocessing (D1), raw dataset (D0), light stemming (D3), and root stemming (D2). Dataset version zero (D0) is finished when creating an excel file. From the first version (D0), three more versions (D2, D1, and D3) were created. This study examines the detection of fake news articles concerning COVID-19 on Facebook with the use of DL approaches, like the Bidirectional Long Short-Term Memory Networks (Bi-LSTM), Bidirectional Encoder Representations from Transformers (BERT) and AraBert of Arabic text and ML techniques Linear Support Vector Machines (SVM) and Random Forest (RF). On testing data-set (D0), BERT yields the greatest accuracy of 97.32%

Keywords: Arabic fake news, Covid-19, deep learning, machine learning.

Corresponding Author:

Raad Sadi Aziz

Faculty of Sciences of Sfax,
University of Sfax, Tunisia.
E-mail: raad1aziz1981@gmail.com

1. Introduction

A claim or piece of information which has been proven to be false is referred to as fake news or rumors [1]. Since it may quickly spread and reach thousands of individuals, false information which is posted on the platforms of social media represents a serious issue. Therefore, with regard to money and time, manual techniques for identifying fake news are not practical. Thus, techniques that could automatically detect fake news are needed in order to stop spreading problematic content and warn public that news that they're reading might not be true. False or misleading COVID-19 information is also becoming a significant issue due to the ongoing COVID-19 pandemic, which has the potential to negatively affect people's health. "Prediction of the possibility that some specific news (news article, editorial, expose, and others) is intentionally false" [2] is the way that fake news detection (FND) had been described. FND tasks have lately sparked a lot of discussion in NLP research community. Scientific community recently became interested in utilizing ML, and specifically DL-based techniques, to uncover these phenomena. Arabic FND is still new and needs a lot of work to catch up to other languages' levels. Thus, a system which automatically aids in confirming the veracity of the shared information concerning COVID-19 pandemic on social media is necessary for the fight against fake news. The task of FND is highly difficult, particularly given the dearth of pandemic-related datasets that are currently available. Using deep/machine learning and NLP methods has been required for an automated FND system [3]. Through the comparison of the text with several well-known corpora which include real as well as fake information, such methods assist in determining whether or not a certain text is fake news [4].

Assigns scores indicating negative or positive opinions to every one of the distinct entities in the text corpus, suggeseted innovative formula for computing the score of polarity for every word that occurs in the text and find it in the negative or positive dictionary [5][6]. Evaluated e-tourism companies utilizing the Iraqi dialect reviews that have been obtained from Facebook. Reviews have been analyzed with the use of the approaches of text mining for the process of sentiment classifications. Sentiment words that have been generated were classified to neutral, positive, and negative comments with the use of the Rough Set Theory [7].

Those systems are discussed in this research in relation to the issue of FND on Facebook throughout COVID-19. Standard Arabic processing is the focus of research on classification and detection of fake news with the use of DL models (BI-LSTM, BERT and ArBert) and ML models (RF, SVM). The suggested model's contribution includes the following: introducing a primary dataset with 64259 words and 5078 sentences obtained from 2539 Arabic-language news posts, both real and fake, that are publicly available. Creates four datasets: pre-processing dataset (D1), raw dataset (D0), light stemming dataset (D3), and root stemming dataset (D2). Those datasets were utilized to test the performance of feature extraction using Bi-LSTM, BERT, ArBert, RF, and SVM, on various datasets.

2- Related work

Despite the fact that the coronavirus pandemic first emerged two years ago, it attracted a lot of interest from the scholarly community. This system will first present a few approaches which were published to stop the spread of COVID-19, after which we will outline the methods that are most pertinent to the goal of detecting fake news. A lot of work has gone into creating sizable COVID-19 datasets since the pandemic started at the end of December 2019. For example, on Contraint@AAAI 2021 Covid-19 FND data-set, [8] Authors compare various supervised text classification techniques. LSTM, convolutional neural networks (CNN), and BERT serve as the foundation for the classification algorithms. Analyze the value of unsupervised learning as it relates to distributed word representations and language model pre-training with the use of unlabeled Covid tweet corpus. On Covid-19 FND dataset, the highest accuracy was 98.41%.

In [9], the researchers use a variety of methods to address the challenges with FND. The statistical distribution related to characters and words throughout the tweets was captured using hand-crafted characteristics the researchers created. This system learned a latent space representation from the collection of word- and character-based n-grams features observed in tweets, possible capturing relevant patterns. It was able to collect contextual information and distinctions between real and fake news of COVID-19 by using a number of BERT-based representations. Learning demonstrated that while it might produce outstanding results for various tasks, like the classification regarding short news, it lags behind other more intricate tasks. To avoid such errors, they built two distinct meta models and took note of what simpler models learned. Through embedding it using a five-layer NN, the 2nd model had learnt a new space from document space representations of the simpler models. On the last (hidden) test set, this new space produced a remarkably accurate representation regarding such problem space with an F1-score of 0.9720.

In [10], to determine whether a news item is fake, the Fake Flow technique models the flow regarding affective information in news. It focuses on news with longer news and depends on the theory that fake news frequently attracts readers' attention through emotional appeals. Flow of affection in news was modeled by authors using neural models, such as CNN and Bidirectional Gated Recurrent Units (Bi-GRU), and those models were tested on three datasets (one data-set that has been created by authors and two available datasets). The scores that have been achieved by Fake Flow are as follows: precision, 0.93; accuracy, 0.96; macro F-1 score, 0.96, and recall, 0.97. They have compared the results with numerous baseline models (such as LSTM, CNN, BERT, HAN, and Long Former). Keep in mind that the long-former model (with macro F-1 score of 0.970) did a little bit better than this one.

In [11], a different method to FND for English involves an ensemble learner technique. According to experimental findings, the ensemble-based approach performs better on FND task than individual learners. With regard to automated classification of news stories, authors employed a ML ensemble approach. This research investigates many textual characteristics which could be utilized to identify between fake and real information. These attributes can be used to train a variety of ML algorithms with the use of different ensemble approaches, and their performance can then be assessed utilizing four real-world datasets. The suggested ensemble learner technique outperforms individual learners, according to experimental evaluation.

In [12], B. Chen et al., 2021 A techniques for optimizing RoBERTa and CT BERT transformer-based language models for FND task was outlined in a prior paper. Here, model robustness was increased through adversarial training. The models have been assessed using a COVID-19 fake news dataset that already existed [11] and compared to cutting-edge techniques. The weighted average F1 score with the highest performance was 99.02%. The results showed higher performances in comparison to different evaluation parameters.

In [13], H. Saadany, 2020, Compared to English FND, Arabic FND is still new, yet it is expanding quickly. For instance, a prior work added 2 new data-sets of the real and fake political news about Middle East. The real news dataset includes 3710 news items from reliable news sources, while the fake news data-set has 3,185 items that have been gathered from 2 Arabic satirical news web-sites. They first conducted exploratory research to determine linguistic characteristics of the Arabic fake news, and they utilized such attributes to build conventional machine language classifiers and neural models in order to specify the item's category. They reported 98.6% accuracy when comparing such methods to a baseline.

Another study that has been carried out by [14] Al-Yahya, 2021, Detection of Arabic Fake News: Comparative Study of NNs and Transformer-Based Methods. The findings of this experiment show that models based on transformers outperform those based on NNs. It was discovered that AraBERT v02 fared better with regard to generalization than all other models that were compared. Even though this effort contributes to the realization of Arabic FND, this system encountered a number of obstacles and constraints. First off, this system employed a small dataset of tweets, and it included data with tweet repetition and unavailable tweets. The data also suffered from noise and unclassified tweets.

In [15], researchers created Arabic fake news using transformers. AraNews, a sizable POS-tagged news dataset that is readily available, was created using this method using actual internet stories and a portion of the speech tagger. Additionally, the authors provided models for the identification of the manipulated Arabic news, and they attained cutting-edge outcomes on Arabic FND assignment with the macro F-1 score of 70.06. Remember that the research's data and models are available for use by anybody.

In [16], Shubha Mishra, 2022, the authors described fundamental concept of the related work in order to present a deep comparative analysis of different literature works. A comparison of various ML and DL methods has been performed for performance evaluation. Which is why, 3 data-sets had been utilized. The comparison was done for many conventional ML and DL methods on 3 liars, fake news, and corpus data-sets. This comparison had concluded that DL methods performance was better than the traditional ML approaches. In this comparison, the Bi-LSTM was able to achieve the optimal rate of detection for fake news and could obtain 95% accuracy and F-1 score.

In [17], Antonio Galli, 2022, had discussed a benchmark analysis of the FND with the use of the classical ML and DL methods (for images as well as texts). DL classifiers have the ability for the automatic extraction of the textual characteristics and analyze the semantic meaning of words based upon the images and context of sentence.

In [18], authors worked a deep NN method that classifies the real and fake news claims through the exploitation of the CNNs. They have utilized Arabic balanced corpus for building their model due to the fact that it unifies the stance rationale, stance detection, relevant document retrieval and fact-checking. This model has been trained on various well selected characteristics. This model outperformed performance of state-of-the-art methods in case of being applied to same Arabic data-set with maximum accuracy of 91%.

In [19], researchers used Evaluation of the DL Methods for the Detection of Covid-19 Fake News. Contraint@AAAI 2021 Covid19 Fake news detection data-set was created using these methods CNN, LSTM, and BERT.They attained cutting-edge outcomes on Arabic FND assignment the optimal accuracy of 98.41% on Covid-19 Fake news detection data-set. Research focuses primarily on the news's content while ignoring other crucial elements such as user characteristics, social network, etc.

In [20], identifying COVID-19 related Fake News via the Neural Stacking.The dataset includes English-language posts from Twitter, Facebook, and Instagram.Research focuses on both learning the handcrafted characteristics of authors and on learning problem space representation using various techniques.The new space had led to a highly convincing representation of that problem space, which had achieved F1-score of 0.972 on nal (hidden) testing set.The research's weak point was that background information should have been added to the representations for producing more in-stance separable representations.

In [21], authors used data-sets of the MultiSourceFake, TruthShades, PoliticalNews, and FakeNewsNet.Through mixing topical and emotive information taken from text, the suggested model, FakeFlow, learns this flow, and they utilized such Results have shown that the FakeFlow achieves better results in comparison with state-of- art approaches.Thereby confirming the significance of capturing flow affective information in news.FakeFlow was trained with a small amount of text, and the results showed that it performs on par with models that require a lot of resources (Longformer and BERT).

In [22], authors worked a FND Using ML Ensemble Methods.They used data -sets utilized Twitter and Facebook, ML models and ensemble methods. Experimental evaluations confirm superior performance of the suggested ensemble learner approach compared with separate learners.There is a lot of unresolved problems in research which should be addressed. For example, understanding crucial components that are involved in the spread of news represents a crucial first step towards the reduction of fake news propagation.

In [23], researchers used Transformer-based Language Model Fine-tuning approaches for COVID-19 FND. Twitter and MicroBlog data-set was created using these methods BERT- transformed-based language models.They Weighted average F1 score had achieved 99.02%. Limited token vocabulary failed as well in getting the whole meaning of the data from a certain domain has been gathered and utilized for downstream fine-tuning.

In [24], authors worked a Fighting an Infodemic: COVID-19 FND. They have utilized Twitter and Facebook. This model outperformed optimal performance of 93.32% F1-score with the SVMs on the testing dataset. More data must be gathered for research, as well as data that have been enhanced by labeling real from fake and collecting data in other languages.

In [25], researchers created Fake or Real? A Study of Arabic Satirical Fake News. Al-Hudood and Al-Ahram Al-Mexici, two Arabic satire news websites, were used to scrape a dataset (3185 news) that was made up entirely of fake information. The 'CNN-Arabic', 'BBC-Arabic', and 'Al-Jazeera news' were used to collect the 3710 news items that make up the real news dataset. Middle Eastern political concerns are addressed in both databases news dataset. Additionally, the authors provided of ML models capable of detecting satirical fake news, and they attained cutting-edge outcomes assignment with the macro Accuracy of up to 98.6%. This study includes an analysis of false negative instances in which satirical news was considered as real. They were the cause of the most classification errors across all models. The most insightful features for the classification task are also explored by this system. Insightful information was offered by both assessments regarding the problem of identifying Arabic satirical fake news.

In [26], Arabic FND: Comparative Study of Neural Networks and Transformer-Based Approaches.The dataset includes ArCOV19-rumors, AraNews, and Arabic news stance (ANS) .Research focuses on NNS and transformer-based language models.Results have shown that transformer-based models perform better than NN-based solutions, increasing the F1 score from 0.83 (optimal NN-based model, GRU) to 0.95 (best transformer-based model, QARiB), and increasing accuracy by 16% in comparison with optimal NN-based solutions. Even though this effort contributes to the realization of Arabic FND, this system encountered a number of obstacles and constraints. First off, this system employed a small dataset of tweets, and it included data with tweet repetition and unavailable tweets. Additionally, the data was hampered by noise and unclassified tweets.

In [27], 2020, researchers used Machine Generation and Detection of Arabic Manipulated and Fake News. AraNews data-set was using these A straightforward approach simply needs a part-of-speech tagger (POS) and the abundance of true stories available online.The first models to detect manipulated Arabic news were developed through research, which also produced the most up-to-date findings on Arabic FND (macro F1 = 70:06). A constraint regarding this work is the dearth of enough data for training the detection models.

Our assessment of the related work in the FND field shows that there has been little work on Arabic FND utilizing neural techniques; hence, more study and research are needed. Furthermore, to the best of our knowledge, DL has not yet been used in research to tackle FND task for Arabic. This work tries to close this knowledge gap and provide some insight into DL for FND task.

2. Methodolgy

The systems used various DL and ML techniques to complete the tasks in the sections that follow. We outline the techniques utilized to create classifier models for each task in this section.

In this research, the suggested FND system's architecture is shown in Figure 1. The initial step in this article is the collection of Facebook data. After consulting with experts in epidemic medicine, this system extracts news that mention COVID-19 topics in the second step, classifies the news manually for determining whether it is real or fake, and then stores the dataset in an excel file for use in our experiments and research. Throughout COVID-19 pandemic, this research aims to create Arabic real and fake news corpora that could be utilized for spotting fake news on the social media. We take the next four steps to meet this need: i) Data collection, ii) rumor/misinformation key-word extractions, iii) division of the datasets into four groups, and iv) classification of fake news. The input characteristics are after that utilized for the training of the various models of machine learning. Ensemble approaches BERT, AraBert and Bi-LSTM feature set employed in this study are the contribution of the suggested method. News is shuffled to achieve an equal distribution of real and fake news in training and test instances. Every data-set has been partitioned to training and testing sets with an 80/20 split, respectively.

In order to achieve accuracy for a certain data-set at the same time as maintaining an ideal balance between bias and variance, the learning algorithms are trained using various parameters. In order to optimize each model for the optimal results, it has been trained many times by utilizing a set of various parameters utilizing a grid search. The power of this system research, presenting a primary dataset contain 5078 sentences and 64259 words acquired from 2539 publicly available fake and real news posts in Arabic language from Facebook. Experiment three of datasets taken from the original group in addition to the original datasets: (D0) raw dataset, (D1) pre-processing dataset, (D2) root stemming dataset, and (D3) light stemming dataset. Our propose used techniques deep learning like BERT, AraBERT, Bi-LSTM and machine learning to purpose comparison such as Random Forest and SVM for the evaluation of performance over the several data-sets. For the evaluation of performance of every one of the models, we have utilized the metrics of accuracy, recall, precision, and F1 score.

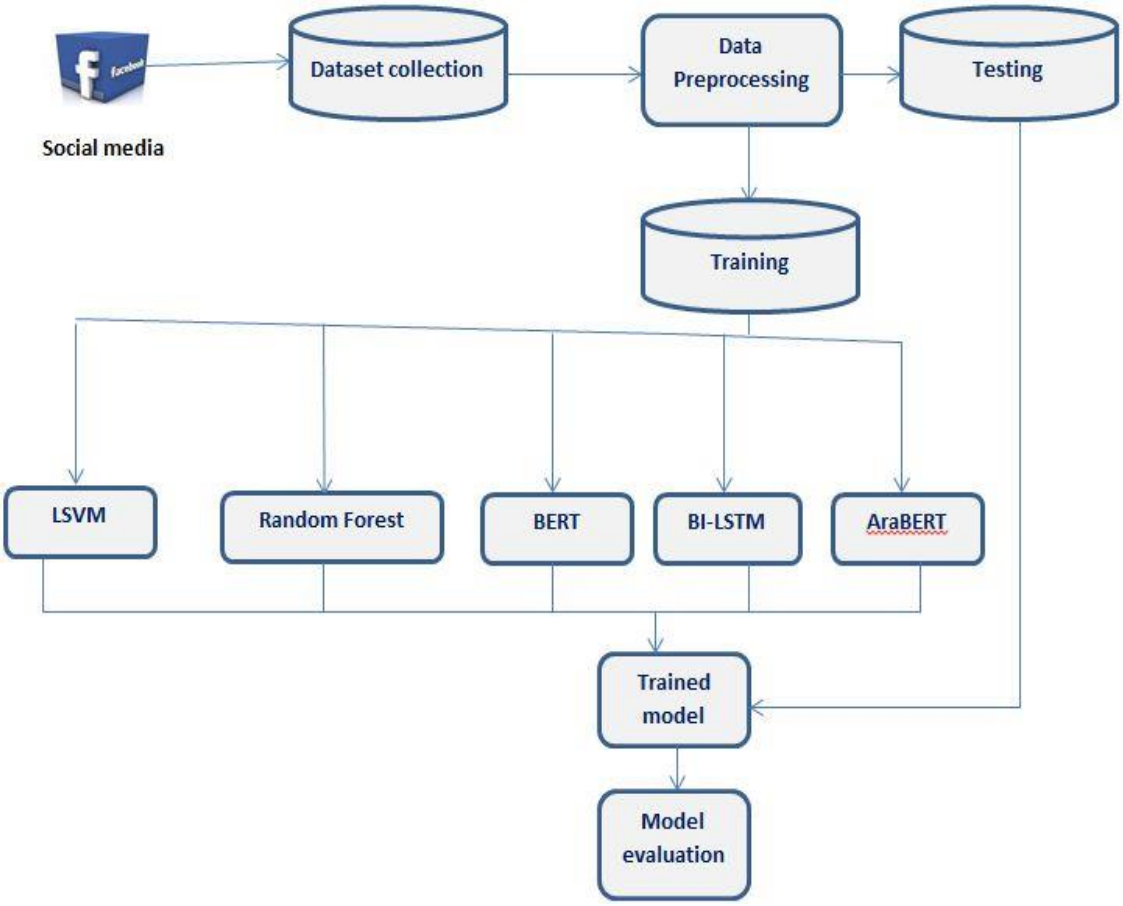


Figure 1. Fake news detection architecture

We have utilized the following learning algorithms in conjunction with this paper method for the evaluation of the performance of the FND classifiers. In the presented section, we present BERT in [28] and go over its specific implementation. Our system consists of 2 steps: fine-tuning and pre-training. The model has been trained on unlabeled data over a variety of the tasks of pre-training throughout pre-training. BERT model has been initialized with pre-trained parameters for the fine-tuning, and all parameters have been adjusted with the use of labeled data from downstream tasks. In spite of being started with same pre-trained parameters, every one of the downstream tasks has its own fine-tuned models. The running example for this part will be the question-answering example in Figure 2. The unified architecture of BERT across all tasks is one of its distinguishing characteristics. The final downstream architecture and the pre-trained architecture barely differ from one another.

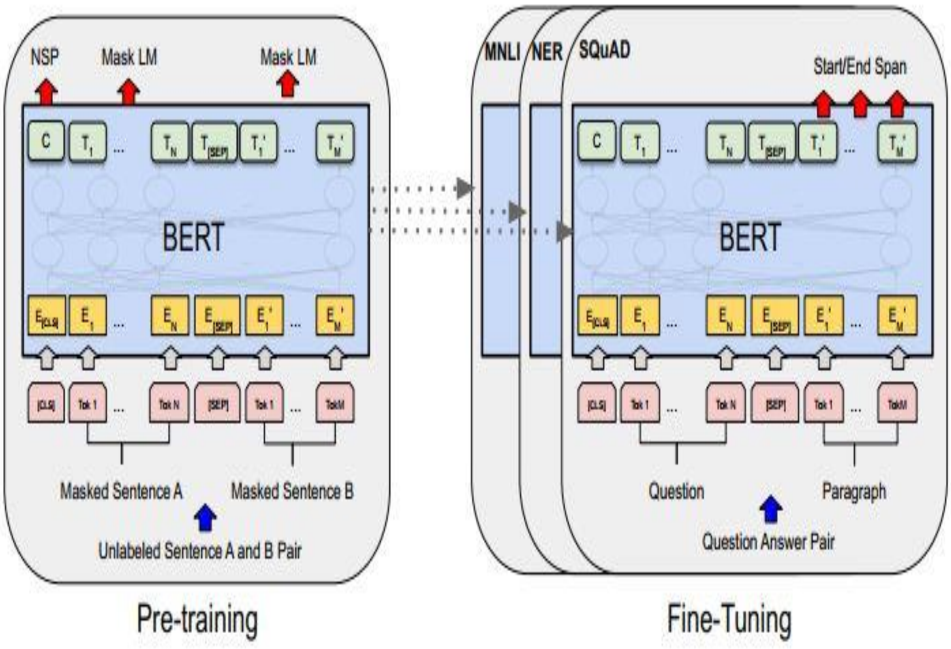


Figure 2. Overall fine-tuning and pre-training procedures for BERT

The same designs have been utilized for the fine-tuning and pre-training, except the output layers. The models have been initialized for a variety of the down-stream activities utilizing the same parameters of pre-trained model. All of the parameters have been adjusted throughout fine-tuning. Each one of the input examples now has a special symbol [CLS] before it, and [SEP] is a special token separating the answers from the questions.

The original implementation that has been described in [29] and made available in tensor-2-tensor library1 serves as the foundation for BERT model, which represents a multi-layer bidirectional Transformer encoder. We will avoid a thorough background description regarding architecture of the model since using Transformers is widespread and our implementation is nearly comparable to original. Rather than that, we will direct readers to [29] and top resources like "Annotated Transformer."

In the present paper, we have denoted the number of layers (in other words, Transformer blocks) as L, hidden size as H, and number of the self-attention heads as A.[3.] We have primarily reported the results on 2 model sizes: **BERT $_{\text{BASE}}$** (L=12, A=12, H=768, Total Parameters=110 M) and **BERT$_{\text{LARGE}}$** (L=24, A=16, H=1024, Total Parameters=340 M). BERT$_{\text{BASE}}$ has been selected in order to have an identical model size to Open AI GPT for the purposes of comparison. Critically, however, BERT Transformer utilizes the bi-directional self-attention, whereas GPT Transformer utilizes the constrained self-attention where each one of the tokens could only attend to the context to its left[4] .

Our used hybrid parameters to obtain on good results in our running, increased numbers of epochs on training, decreased batch size, cross validation on 20% of dataset as validation data.

For various NLP classification and language understanding tasks, deep contextualized language models, like BERT [30], have lately produced significant advances. We improved AraBERT (V1) for the classification task in this work [31]. The encoder for BERT has 12 self-attention heads, 12 Transformer blocks, and one hidden size of 768. This architecture is the same as that used for AraBERT. It is trained on a sizable Arabic news corpus of 2539 news, of which 1387 (54.62%) are real and 1158 (45.60%) are fake. We utilized the PyTorch3 implementation from HuggingFace4 as it includes pre-trained weights and vocabularies. AraBERT was found to produce better results compared to multi-lingual BERT from Google, which has been trained upon Arabic Wikipedia alone [15]. The method is indicated as Bi-LSTM in literature. [32] Also employed the Bi-LSTM and we have applied this same method with various sets of feature. Figure 3 presents the details of the Bi-LSTM.

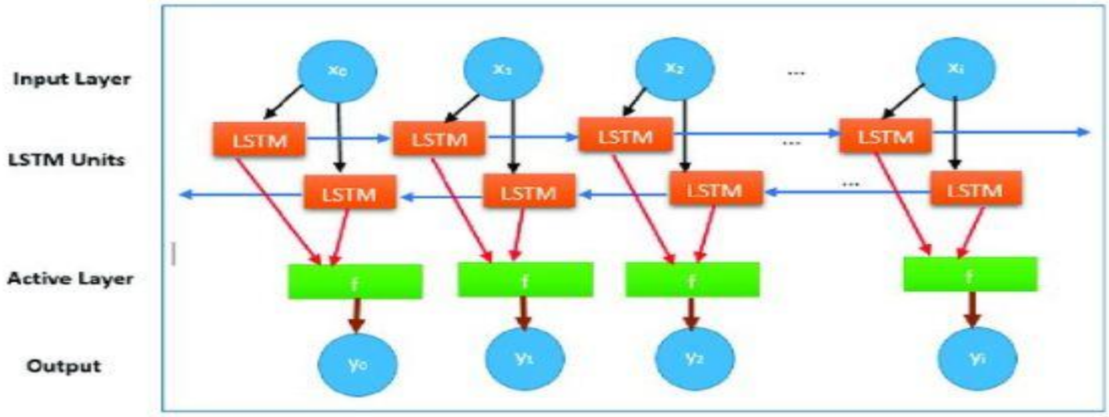


Figure 3. Structure of bi-directional LSTM (biLSTM) algorithm

DT, a supervised learning model, is simplified in RF. In RF, a substantial number of DTs estimate every class's outcome independently, with the final prediction depending on the class that is receiving the most votes. RF has a lower error rate compared to other models since there is less correlation between the trees [33]. Our RF model was trained with an accuracy of parameters, in other words, varied numbers of the estimators have been utilized in the grid search, to discover optimal model which could properly predict results. Various approaches could be used to find a split in a DT depending on the classification or regression issue. We used Gini index as cost function for estimating a split in data-set for classification problem. The Gini index has been estimated through the subtraction of the total of squared likelihoods of every class from one. The following mathematical equation is used for calculating the Gini index (G_{ind}) [34]:

$$G_{ind}= 1 - \sum_{i=1}^{c}(Pi) \qquad (1)$$

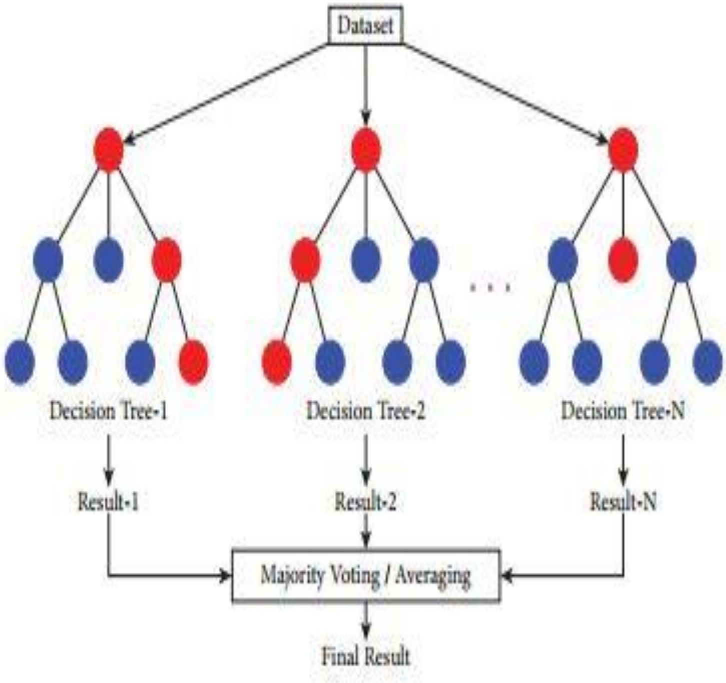


Figure 4. Illustration of random forest trees

The linear SVM method used in this system was suggested in [35]. We have trained linear SVM on feature set with 5-fold cross validation to achieve a meaningful comparison in the text. Shown in Figure Overview of Support Vector Machines

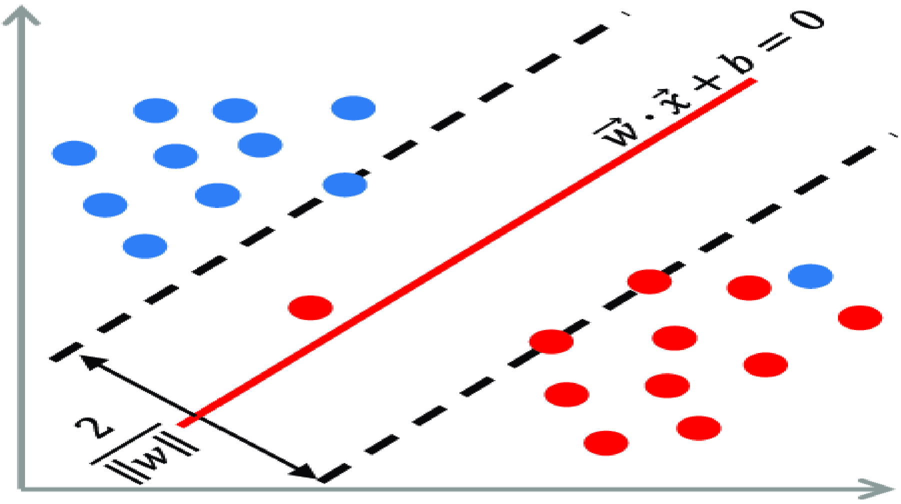


Figure 5. Support vector machine

4. Data collection

The techniques utilized to transform raw data into excel files, the sources from which the dataset was compiled, and the procedure for gathering data for two different classes. The original dataset is created by converting raw data, which is also known as constructing raw data. The pre-processing dataset, D1, is created due to the pre-processing (D0). D1 is subjected to light stemming and root stemming techniques to produce D3 and D2. The outcomes of using such techniques will be examined statistically. Resources that have been utilized for collecting the data-set and reasoning behind the selection of such resources Corona virus at the start of 2020s yet, other individuals utilized social media sites like Facebook to mobilize crowds and to enthusiastically support the corona. People were able to express their opinions more openly thanks to the internet's accessibility. The collect dataset operation is shown in Figure 6.

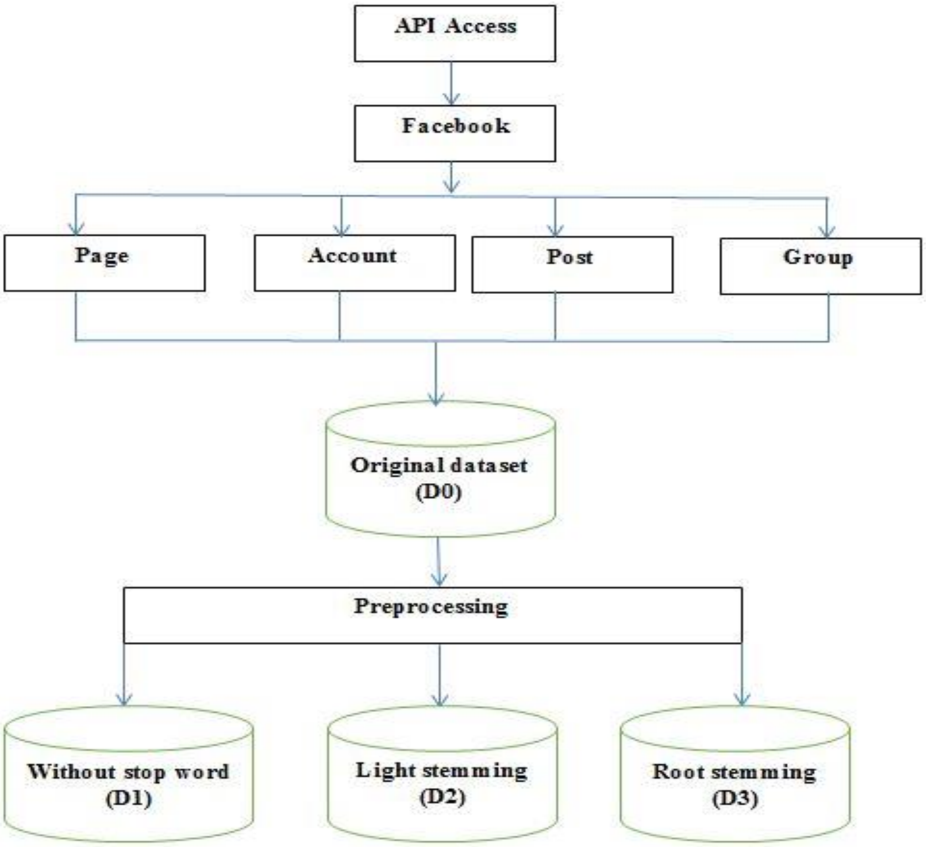


Figure 6. Data collection

Finding fake Arabic news online poses a significant challenge. This difficulty includes a lack of some common Arabic data-set. Therefore, the data-set that had been gathered from diverse resources has been generated for this experimental study. The selected data-sets came from various sources, including (Facebook). The datasets were divided into three groups (non-fake and fake). Table 1 provides a summary of such groups.

Table1. Number of selected news

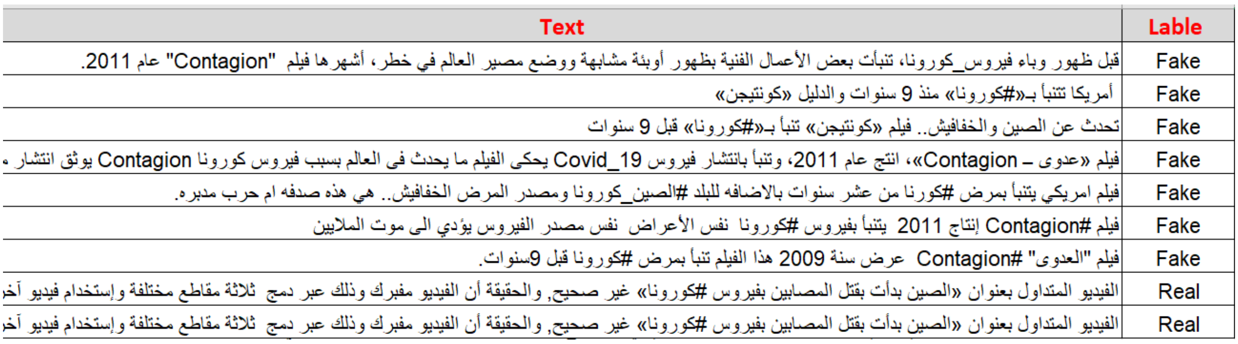
Label of news	news number)	Arabic meaning
Fake	1158 (45.60%)	كاذبة
Real	1387 (54.62%)	غير كاذبة
Total	2539 (100%)	

Fake news is any news that had been written by writers or authors using fake information. The dataset must be compiled from different Arabic authors who published news in different categories (fake and real). There are three different sorts of news (medium, short, and large) based on the news size, as indicated in Table2.

Table 2. Number of news

News	No. of news	No. of words
short news	450	≤ 150
Medium news	1350	≤ 250
long news	739	>250
Total	2539	-

The collection of news only includes news written in contemporary standard Arabic; all vernacular Arabic news have been omitted. It is required to omit anything that is not written in Modern Standard Arabic because the data collected is not limited to a single nation or dialect. To put the news in its folder, it is manually divided into two categories based on labels (real or fake). Each news item is distinct and retained in its raw form; as a result, such news items have not undergone any stemming, cleaning, or other pre-processing.
Classify news into two categories according to labels (fake and real). Figure 4 shows an example of fake news and real.



Text	Lable
قبل ظهور وباء فيروس_كورونا، تنبأت بعض الأعمال الفنية بظهور أوبئة مشابهة ووضع مصير العالم في خطر، أشهرها فيلم "Contagion" عام 2011.	Fake
أمريكا تتنبأ بـ«#كورونا» منذ 9 سنوات والدليل «كونتيجن»	Fake
تحدث عن الصين والخفافيش.. فيلم «كونتيجن» تنبأ بـ«#كورونا» قبل 9 سنوات	Fake
فيلم «عدوى – Contagion»، انتج عام 2011، وتنبأ بانتشار فيروس Covid_19 يحكى الفيلم ما يحدث فى العالم بسبب فيروس كورونا Contagion يوثق انتشار	Fake
فيلم امريكي يتنبأ بمرض #كورنا من عشر سنوات بالاضافه للبلد #الصين_كورونا ومصدر المرض الخفافيش.. هي هذه صدفه ام حرب مدبره.	Fake
فيلم #Contagion إنتاج 2011 يتنبأ بفيروس #كورونا نفس الأعراض نفس مصدر الفيروس يؤدي الى موت الملايين	Fake
فيلم "العدوى" #Contagion عرض سنة 2009 هذا الفيلم تنبأ بمرض #كورونا قبل 9سنوات.	Fake
الفيديو المتداول بعنوان «الصين بدأت بقتل المصابين بفيروس #كورونا» غير صحيح, والحقيقة أن الفيديو مفبرك وذلك عبر دمج ثلاثة مقاطع مختلفة وإستخدام فيديو آخ	Real
الفيديو المتداول بعنوان «الصين بدأت بقتل المصابين بفيروس #كورونا» غير صحيح, والحقيقة أن الفيديو مفبرك وذلك عبر دمج ثلاثة مقاطع مختلفة وإستخدام فيديو آخ	Real

Figure7. Example of a fake and real news

Four versions of the datasets have been utilized in this study (D1, D0, D3, and D2). To understand the effects of ArBERT, BERT, RF, LSTM, and SVM algorithms on all datasets, four datasets were created. Additionally, the research analyzes them with regard to DL and ML to determine the efficacy of the preprocessing (D1), raw dataset (D0), light stemming (D3), and root stemming (D2). Dataset version zero (D0) is finished when creating an excel file. According to Table 3, three new versions (D2, D1, and D3) were created by deriving from the first version (D0).

Table 3: Dataset versions

Version	Dataset name	Description
D0	Raw-data	Original data
D1	Preprocessing	This version has been built though the application of preprocessing steps on the raw data
D2	Root stemming	Built through applying ISRI root stemmer on D1
D3	Light stemming	This version has been built through the application of the light stemmer on D1

4.1. Preprocessing (Dataset D1)

HTML tags, Scripts, punctuation, links, and other noise data are often present in the online news (i.e. the fake news). The extraction regarding influential words from fake news could be challenging as a result of such noisy data's lack of the meaning. Thus, a pre-processing approach has to be utilized for removing this data since maintaining them could make the process of the classification more challenging due to the fact that each data can be evaluated as a single dimension. Tokenization, punctuation removal, removing unnecessary words, normalization, and removing stop words are preprocessing steps.

The news was split up into a sequence of words in this step.Splitting the news into words through the tokenization process. Algorithm shows the steps of tokenization.

```
Algorithm  : Tokenization
  Input: Texts $T$
$\boldsymbol{Tok} = \{\ \}$, list of tokens
For $\boldsymbol{T}$ in dataset
    split $T$ by (space)
    $\boldsymbol{Tok}$.appeand ($\boldsymbol{T_i}$), split the texts into tokens
EndFor
Output:
$\boldsymbol{Tok}$, list of tokens (words)
```

Punctuation marks make the line of coherence connecting sentences, paragraphs, and phrases easier to understand and read. The Arabic language uses a variety of punctuation marks, including apostrophes, commas, quotes, and question marks. It is required to remove any punctuation marks (e.g., remove it) after the tokenization procedure has divided the news into words using white spaces such as (:".،?] //…[/*'{}؟' - _ ;).

Various unneeded words, such as English or Arabic numbers, non-Arabic words or characters, and others which aren’t in punctuation marks, are eliminated in this step. Many regular expressions are available to accomplish this, as seen in Table 4.

Table 4. List of regular expressions

Regular Expressions	Results
[a-zA-Z]+	Eliminate English characters
[0-9]+	Eliminate Arabic numbers
[0-9]+	Eliminate English numbers
#$%|@^~(&*)+	Eliminate other symbols

The process of transforming the textual content to one canonical framework is known as Arabic news normalization. Put differently, by normalizing the news prior to processing it, it is possible to separate concerns because input is certain to be consistent before any operations are performed on it. Since Arabic words contain a variety of shapes, which results in low accuracy and high dimensions, the D1 dataset in this work is normalized and transformed to unified form. Three steps make up the normalization process: removal of the diacritics, removal of tatweel, and normalization of the letters. This process begins by the removal of all of the diacritics from Arabic words and transformation of an alphabetic word to another. Figure5 depicts diacritics which may be eliminated. In the case where this process isn’t applied, there would be too many words, which will result in increasing the search space as well as required time. For instance, (" تعبيراً") become ("تعبيرا"), in the case where the diacritic ("ً") is removed from words, the result will be one word.

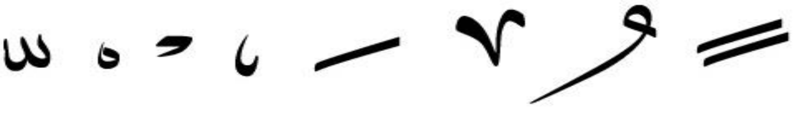


Figure 8. Arabic diacritics

Tatweel in Arabic represents a character representing the elongation (——), whicih is referred to as ("تطويل") as well. In other words, it can be described as the increase of the length of a line of the news through the expansion of the spaces between the words or the separate letters. For instance, "كارونــــــــــــا", "كارونـــــا" and "كارونـــــا" is the same word as words "كارونا". The last step of normalization requires making a distinctive letter for some of the letters. In Arabic language, there are numerous ways for writing the characters, which had been characterized by Table5.

Table 5. Latter arabic language normalization

Original	Become	Examples	
إأآا	ا	أعراض	اعراض
		symptoms	Symptoms
ى	ي	قوى	قوي
		strong	Strong
ؤ	ء	اداؤه	اداءه
		performance	Performance
ئ	ء	يتشائم	يتشاءم
		pessimistic	Pessimistic
ة	ه	صحيفة	صحيفه
		newspaper	newspaper

After being compared with the list of the Arabic stop words, an Arabic stop words could be eliminated from a news article. In the present investigation, 2 lists of stop words have been utilized. The first list was created with the use of NLTK library in Python, and the second list contains words that are irrelevant to the sentence as it was shown in table6 for example.

Table 6. Shown stop word

word	Become
وماذا	و-ماذا
فغادر	ف-غادر
فحسبه	ف-حسب-ه
وبغيري	و-ب-غير-ي
لهم	لهم

4.2. Root stemming (dataset D2)

The process of separating words from their roots has been referred to as ISRI stem. The data-set version D2 in this study is subjected to the root stemmer. There are various steps in the suggested stemmer.

4.3. Light stemming (dataset D3)

Due to its extensive vocabulary, extensive storage of synonyms, and numerous grammar rules, the Arabic language is extremely unique with regard to its stemming. Due to this, there must be specific rules on how to handle the word additions (suffixes, prefixes). One of the most crucial steps in creating dataset D3 is this one. It eliminates linguistic additions like suffixes and prefixes. All words are derived from a single origin in this fashion since they are represented in their original form. Therefore, this stemmer's implementation results in a decrease in both space and time, increasing its efficiency in handling all rules. By changing the ISRI stemmer, which represents a number of the criteria that determines the way of applying stemming on a particular word, an approach for the Arabic light stemmer was presented. The Arabic light stemmer's rules for deleting waw ("و"), prefixes, and suffixes. Table 7 illustrates the rules of waw, suffixes, and prefixes that were built. Those rules, which were gathered from ISRI, are just written in the manner described above; they differ significantly from original algorithms, which involved numerous phases.

Table 7. Prefix, suffix and waw

Method	Word Length	Letter Removed from the word	Letters
Waw	w >= 4	1	و
Prefix	p>=6	3	كال, بال, ولل, وال
	p>=5	2	لل, ال
	s>=6	3	تمل, همل, تان, كمل, تين
Suffix	s>=5	2	ون, ات, ان, ين, تن, كم, هن, نا, يا, ها, تم, كن, ني, وا, ما,هم

5. Performance metrics

This system employed many metrics for the assessment of algorithm effectiveness. Confusion matrix has been considered as a foundation of most of them. Confusion matrix, including the 4 parameters false positive (FP), true positive (TP), false negative (FN), and true negative (TN), is a tabular representation regarding a classification model's efficiency on testing dataset.

Table 8. Confusion matrix

	Predicted real	Predicted fake
Actual real	Real positive (RP)	Fake negative (FN)
Actual fake	Fake positive (FP)	Real negative (RN)

Table 9. Confusion matrix for fake news and real news

	Fake	Real
Fake	159	11
Real	17	167

5.1. Accuracy

The most common metric for percentage of correctly predicted observations—whether false or true —is accuracy. The next equation could be utilized for the determination of the accuracy of a model's performance:

$$Accuracy = \frac{\mathrm{RP + RN}}{\mathrm{RP + RN + FP + FN}} \tag{2}$$

The accuracy of 92.02% indicate a good model, yet because such system is training a classification model in this case, a news item that was predicte d as real but has actually been false (false positive) may have unfavorable effects. Likewise, if a news item was predicted as false but contained factual information, this may result in causing trust issues. Thus, our system has employed 3 additional metrics, which are—recall, precision, and F1-score—that are responsible for incorrectly classified observation.

5.2. Recall

The total number of positive classifications out of the real class has been referred to as the recall. In our case, it represents proportion of all of the real news that has been predicted to be real Recall of 90.34%.

$$Recall = \frac{\mathrm{RP}}{RP+FN} \tag{3}$$

5.3. Precision

Ratio of the true positives to all events that have been anticipated as real is what the precision of 93.52%, however, indicates. In our case, precision is the proportion of the positively predicted (i.e. real) news that is designated as real:

$$\mathbf{Precision} = \frac{\mathrm{RP}}{RP+FP} \tag{4}$$

5.4. F1-Score

Precision/recall trade-off has been characterized by F1-score. It specifies harmonic average of every one of the pairs. Which is why, it considers fake negative as well as fake positive observations. The formula below could be utilized for the determination of an F1-score:

$$\mathrm{F1-score} = \frac{2(\mathrm{Precision} \times \mathrm{Recall})}{Precision+ Recall} \tag{5}$$

6. Results and discussion

The techniques utilized to transform raw data into excel files, the sources from which the dataset was compiled, and the procedure for gathering data for two different classes. Building raw data results in the creation of raw data, often known as the original dataset D0. The pre-processing dataset, D1, is created due to pre-processing D0. D1 is subjected to light stemming and root stemming techniques to produce D3 and D2. The outcomes of using such techniques were statistically analyzed. There is 2539 news that has been pulled from Facebook. Datasets were divided into two groups (real news and fake news). Those categories contain 1158 (45.60%) fake news and 1387 (54.62%) real news. There are three categories of news: medium (1350), short (450), and large (2539), according to the magnitude of the news. The accuracy of each algorithm on the five datasets under consideration is summarized in Table 10. It is clear that BERT algorithm attained a maximum accuracy of 97.32 % on D0 (original datasets). On D0 (original datasets), AraBERT had a 95.28% accuracy. On D1 (preprocessing datasets), the accuracy of the linear SVM, bagging classifiers, multilayer perceptron, and boosting classifiers was 95%. RF's high accuracy on D2 (light stemming datasets) and D3 is 91%. (Root stemming datasets). Bi-LSTM underperformed all other benchmark algorithms in algorithms of performance. Bagging BERT is the top performing algorithm across all datasets. Surprisingly, RF and linear BI-LSTM did poorly on D1 and D0. For D1 and D0 accuracy, 89% for RF and 87.57% for BI-LSTM show a comparable pattern. The best algorithm on (D0, D1, D2, and D3 taken together) is BERT (97.32% accuracy). The algorithm that performs the wors, BI-LSTM, achieved an accuracy of 87.57% on (D0, D1, and D3).

Table 10. Overall accuracy score for each dataset

	D0	D1	D2	D3
BERT	97.32 %	95.98%	95.5%	94.20%
AraBERT	95.28%	94.49%	93.70%	93.44%
BI-LSTM	87.57%	87.57%	90.20 %	87.57%
Random Fores	89%	89 %	91%	91%
SVM	93%	95%	94%	93%

The accuracy related to all algorithms across the 4 datasets is summarized in Figure 9. BERT, which performs best overall (accuracy: 97.32%), outperforms BI-LSTM, which performs lowest overall (accuracy: 87.57%). Yet, accuracy score alone is not a reliable indicator of a model's performance, thus we additionally assess learning models' performance using precision, recall, and F1-score.

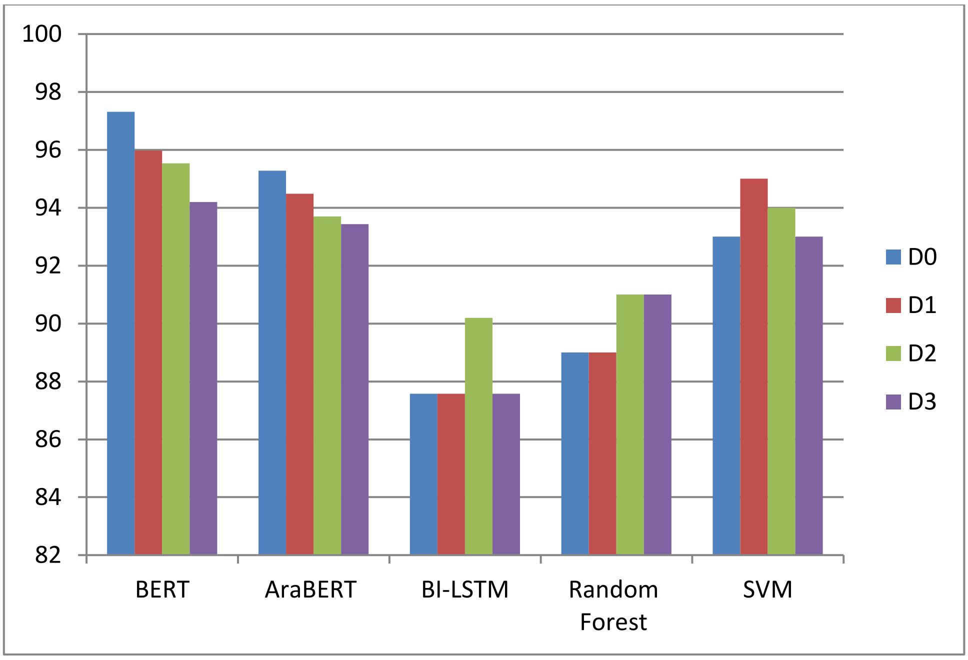


Figure 9. Accuracy over all datasets

Tables 11–13 list each algorithm's precision, recall, and F1 score for the 4 datasets. Table 11 shows that boosting SVM produced the best results with regard to precision. On (D1) datasets, boosting SVM has a precision of 95%. On all datasets, AraBERT attained a precision of 92%. On datasets (D2, D3), RF attained a precision of 91%, and on the datasets (D0, D1), 88%. The corresponding score regarding BERT is 78% on data-sets (D1, D2, and D3) and 11% on datasets (D0). By earning a recall score of 97% on the data-sets (D1, D2, and D3) and 96% on the datasets (D0), SVM stands best according to the recall performance criteria.The recall for boosting RF, which obtained 97% on datasets (D1), and 96% on datasets (D0, D2, D3), are both quite similar.A recall score of 94% on the datasets (D1, D2), 93% on the datasets (D0), and 86% on the datasets (D3) make BI-LSTM the best performing benchmark method among the algorithms tested. On all datasets, AraBERT achieved a recall of 92%. The algorithms performed similarly on F1- score and in terms of precision. Boosting SVM earned the maximum F1-score of 96% of all approaches, AraBERT achieved F1-score of 92% on all datasets, succeeded by bagging RF and BI-LSTM.

The average performance of the learning algorithms on all of the data-sets utilizing recall, precision, and F1-score is shown graphically in Figure 10. As can be observed, there are not many differences in how well learning algorithms work when used with different performance metrics.

Table 11. Precision on 4 datasets

	D0	D1	D2	D3
BERT	11%	78%	78%	78%
AraBERT	92%	92%	92%	92%
BI-LSTM	87%	87%	90%	89%
Random Fores	88%	88%	91%	91%
SVM	93%	95%	94%	89%

Table 12. Recall on 4 datasets

	D0	D1	D2	D3
BERT	33%	35%	34%	48%
AraBERT	92%	92%	92%	92%
BI-LSTM	93%	94%	94%	86%
Random Fores	96%	97%	96%	96%
SVM	96%	97%	97%	97%

Table 13: F1-score on the 4 datasets

	D0	D1	D2	D3
BERT	17%	19%	17%	44%
AraBERT	92%	92%	92%	92%
BI-LSTM	90%	90%	92%	87%
Random Fores	92%	92%	93%	93%
SVM	95%	96%	95%	97%

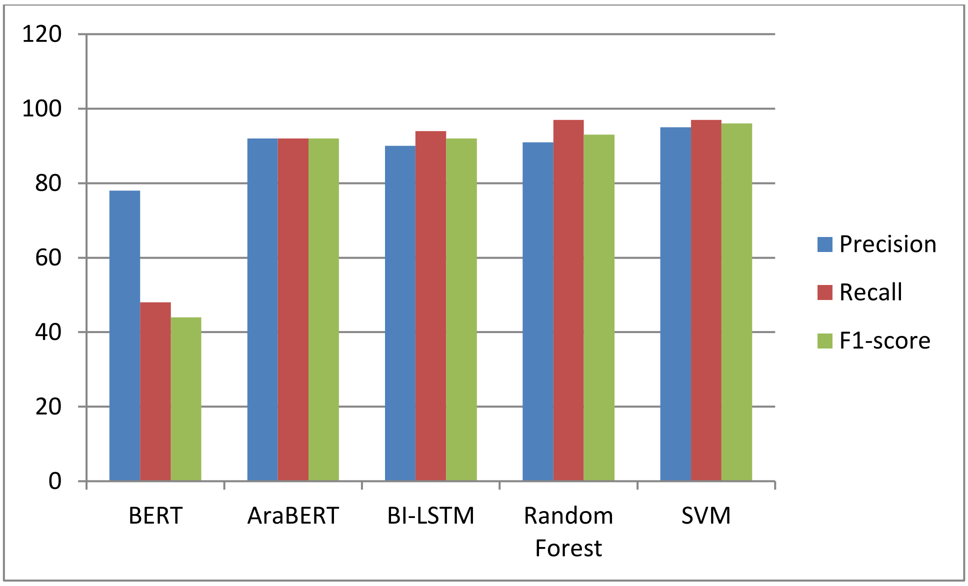


Figure 10. Recall, precision, and F1-score over all datasets

Compared to other learning models, the ensemble SVM outperformed them in terms of recall, F1-score and precision. The working principle regarding SVM that effectively finds flaws and minimizes them each time, is the primary factor that contributes to its exceptional performance. As a result, the model is able to correctly detect misclassified points in each time, while regularization parameters have been employed in order to lessen overfitting problem. There are a few possible causes for the high accuracy: first, the BERT model is optimized via a thorough grid search with several hyper parameters; second, a few datasets contain authors with comparable writing styles, which contributed to BERT model's 97% accuracy.

7. Conclusion

Manually classifying the news involves thorough knowledge regarding the field and the capability to spot text anomalies. In the present paper, The studies that have been reviewed where selected based upon their relationship with this study. In which the limitations in the related literature are stated. we have examined the issue of classifying fake news utilizing ensemble methods, DL, and ML models. The information we utilized in our research was gathered from Facebook's corona virus-specific news data on WWW. The research's main goal is finding textual patterns distinguishing between real and fake news. For the purpose of getting the highest level of accuracy, the learning models underwent training and parameter tuning. We compared results of each one of the algorithms using a variety of the indicators of performance. In comparison to other algorthms, the ensemble BERT has consistently outperformed them in terms of performance accuracy across all datasets. the best results of its great for categorized data accuracy of (97.32%) .There are numerous unresolved problems with the FND, which should be researched. For example, understanding crucial components that are involved in spreading news is a crucial first step towards the reduction of fake news propagation. The key sources that are engaged in the spread of fake news could be identified using ML approaches. Another prospective future direction is real time fake news detection in videos, i. e. images, text and voice.

Declaration of competing interest

The authors declare that they have no any known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

References

[1] P. Patwa, M. Bhardwaj, and V. Guptha, “Overview of CONSTRAINT 2021 shared tasks: detecting English COVID- 19 fake news and hindi hostile posts,” in *Proceedings of the CONSTRAINT 2021*, Delhi, India, 2021.

[2] N. K. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news: Automatic Deception Detection: Methods for Finding Fake News,” *Proc. Assoc. Inf. Sci. Technol.*, vol. 52, no. 1, pp. 1–4, 2015.

[3] R. Oshikawa, J. Qian, and W. Y. Wang, “A survey on Natural Language Processing for fake news detection,” *arXiv [cs.CL]*, 2018.

[4] M. K. Elhadad, K. Fun Li, and F. Gebali, “Fake news detection on social media: A systematic survey,” in *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, 2019.

[5] D. H. Abd, A. R. Abbas, and A. T. Sadiq, “Twitter Sentiment Analysis and Events Prediction Using Lexicon Sentiment Analysis for Iraqi Vernacular,” *Bull. Electr. Eng. Informatics*, vol. 10, pp. 283–289, 2021.

[6] A. Khalid Al-Mashhadany, A. T. Sadiq, S. Mazin Ali, and A. Abbas Ahmed, “Healthcare assessment for beauty centers using hybrid sentiment analysis,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 2, p. 890, 2022.

[7] N. F. AL-Bakri, J. F. Yonan, and A. T. Sadiq, “Tourism companies assessment via social media using sentiment analysis,” *Baghdad Sci. J.*, vol. 19, no. 2, p. 0422, 2022.

[8] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, “Evaluating deep learning approaches for Covid19 fake news detection,” in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, Cham: Springer International Publishing, 2021, pp. 153–163.

[9] B. Koloski, T. Stepišnik-Perdih, S. Pollak, and B. Škrlj, “Identification of COVID-19 related fake news via neural stacking,” in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, Cham: Springer International Publishing, 2021, pp. 177–188.

[10] B. Ghanem, S. P. Ponzetto, P. Rosso, and F. Rangel, “FakeFlow: Fake news detection by modeling the flow of affective information,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.

[11] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake news detection using machine learning ensemble methods,” *Complexity*, vol. 2020, 2020.

[12] B. Chen *et al.*, “Transformer-based language model fine-tuning methods for COVID-19 fake news detection,” *arXiv [cs.CL]*, 2021.

[13] P. Patwa *et al.*, “Fighting an Infodemic: COVID-19 Fake News Dataset,” in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, Cham: Springer International Publishing, 2021, pp. 21–29.

[14] H. Saadany, E. Mohamed, and C. Orasan, “Fake or real? A study of Arabic satirical fake news,” *arXiv [cs.CL]*, 2020.

[15] M. Al-Yahya, H. Al-Khalifa, H. Al-Baity, D. AlSaeed, and A. Essam, “Arabic fake news detection: Comparative study of neural networks and transformer-based approaches,” *Complexity*, vol. 2021, pp. 1–10, 2021.

[16] E. M. B. Nagoudi, A. Elmadany, M. Abdul-Mageed, T. Alhindi, and H. Cavusoglu, “Machine generation and detection of Arabic manipulated and fake news,” *arXiv [cs.CL]*, 2020.

[17] S. Mishra, P. Shukla, and R. Agarwal, “Analyzing machine learning enabled fake news detection techniques for diversified datasets,” *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–18, 2022.

[18] A. Galli, “A comprehensive Benchmark for fake news detection,” *Journal of Intelligent Information Systems*, vol. 59, no. 1, pp. 237–261, 2022.

[19] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, “Evaluating deep learning approaches for covid19 fake news detection,” in *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event*, Springer International Publishing, 2021, pp. 153–163.

[20] B. Koloski, T. Stepišnik-Perdih, S. Pollak, and B. Škrlj, *Identification of COVID-19 related fake news via neural stacking. In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event*. Springer International Publishing, 2021.

[21] B. Ghanem, S. P. Ponzetto, P. Rosso, and F. Rangel, “FakeFlow: Fake news detection by modeling the flow of affective information,” *arXiv [cs.CL]*, 2021.

[22] S. Elyassami, S. Alseiari, M. ALZaabi, A. Hashem, and N. Aljahoori, “Fake news detection using ensemble learning and machine learning algorithms,” in *Studies in Computational Intelligence*, Cham: Springer International Publishing, 2022, pp. 149–162.

[23] B. Chen *et al.*, “Transformer-based language model fine-tuning methods for COVID-19 fake news detection,” in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, Cham: Springer International Publishing, 2021, pp. 83–92.

[24] P. Patwa *et al.*, “Fighting an Infodemic: COVID-19 Fake News Dataset,” *arXiv [cs.CL]*, 2020.

[25] H. Saadany, C. Orasan, and E. Mohamed, “Fake or real? A study of Arabic satirical fake news,” in *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, 2020, pp. 70–80.

[26] D. S. Abd Elminaam, A. Abdelaziz, G. Essam, and S. E. Mohamed, “AraFake: A deep learning approach for Arabic fake news detection,” in *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 2023.

[27] E. M. B. Nagoudi, A. Elmadany, M. Abdul-Mageed, T. Alhindi, and H. Cavusoglu, “Machine generation and detection of Arabic manipulated and fake news,” *arXiv [cs.CL]*, pp. 69–84, 2020.

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

[29] D. Meenakshi and A. Rahim Mohamed Shanavas, “Transformer induced enhanced feature engineering for contextual similarity detection in text,” *Bull. Electr. Eng. Inform.*, vol. 11, no. 4, pp. 2124–2130, 2022.

[30] W. Antoun, F. Baly, and H. Hajj, “AraBERT: Transformer-based model for Arabic language understanding,” *arXiv [cs.CL]*, 2020.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional Transformers for language understanding,” *arXiv [cs.CL]*, 2018.

[32] P. Bhattacharya, S. B. Patel, R. Gupta, S. Tanwar, and J. J. P. C. Rodrigues, “SaTYa: Trusted bi-LSTM-based fake news classification scheme for smart community,” *IEEE Trans. Comput. Soc. Syst.*, vol. 9, no. 6, pp. 1758–1767, 2022.

[33] M. Y. Khan, A. Qayoom, M. S. Nizami, M. S. Siddiqui, S. Wasi, and S. M. K.-U.-R. Raazi, “Automated prediction of Good Dictionary EXamples (GDEX): A comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques,” *Complexity*, vol. 2021, pp. 1–18, 2021.

[34] J. Van Ryzin, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and Regression Trees,” *J. Am. Stat. Assoc.*, vol. 81, no. 393, p. 253, 1986.

[35] J. L. Balcázar, Y. Dai, J. Tanaka, and O. Watanabe, “Provably fast training algorithms for support vector machines,” *Theory Comput. Syst.*, vol. 42, no. 4, pp. 568–595, 2008.