



Developing a fake news identification model with advanced deep language transformers for Turkish COVID-19 misinformation data

Mehmet BOZUYLA^{1,*} , Akın ÖZÇİFT² 

¹Department of Electrical-Electronics Engineering, Faculty of Engineering, Pamukkale University, Denizli, Turkey

²Department of Software Engineering, Hasan Ferdi Turgutlu Technology Faculty, Manisa Celal Bayar University, Manisa, Turkey

Received: 11.06.2021

Accepted/Published Online: 31.10.2021

Final Version: 21.03.2022

Abstract: The massive use of social media causes rapid information dissemination that amplifies harmful messages such as fake news. Fake-news is misleading information presented as factual news that is generally used to manipulate public opinion. In particular, fake news related to COVID-19 is defined as ‘infodemic’ by World Health Organization. An infodemic is a misleading information that causes confusion which may harm health. There is a high volume of misinformation about COVID-19 that causes panic and high stress. Therefore, the importance of development of COVID-19 related fake news identification model is clear and it is particularly important for Turkish language from COVID-19 fake news identification point of view. In this article, we propose an advanced deep language transformer model to identify the truth of Turkish COVID-19 news from social media. For this aim, we first generated Turkish COVID-19 news from various sources as a benchmark dataset. Then we utilized five conventional machine learning algorithms (i.e. Naive Bayes, Random Forest, K-Nearest Neighbor, Support Vector Machine, Logistic Regression) on top of several language preprocessing tasks. As a next step, we used novel deep learning algorithms such as Long Short-Term Memory, Bi-directional Long-Short-Term-Memory, Convolutional Neural Networks, Gated Recurrent Unit and Bi-directional Gated Recurrent Unit. For further evaluation, we made use of deep learning based language transformers, i.e. Bi-directional Encoder Representations from Transformers and its variations, to improve efficiency of the proposed approach. From the obtained results, we observed that neural transformers, in particular Turkish dedicated transformer BerTURK, is able to identify COVID-19 fake news in 98.5% accuracy.

Key words: Infodemic, fake news, BerTURK, language transformers, machine learning, COVID-19

1. Introduction

Fake news is defined as “particular news articles that originate either on mainstream media (online or offline) or social media and have no factual basis, but are presented as facts” [1]. Though the fake news concept is not new, it is becoming a health issue with its dissemination through social media. The popular use of social media channels spread ideas or news through shares, likes or retweets; hence social media users are inevitably exposed to various types of unfiltered information that is suspicious to be genuine or not [2]. In this context, fake news can easily effect public opinion with possible unhealthy consequences such as COVID-19 influences.

The circulation of false COVID-19 information through social media, i.e fake COVID-19 news, has harmfully influenced many people while increasing panic and stress during the epidemic. This is defined as infodemic by World Health Organization (WHO) director and he underlined the infodemy risk as “We are not

*Correspondence: mbozuyula05@posta.pau.edu.tr

just fighting an epidemic; we are fighting an infodemic. Fake news spreads faster and more easily than this virus, and is just as dangerous” [3].

There are some specific consequences of underlined infodemy progress: i) acceleration of virus epidemy, ii) growth of antivaccination community and iii) threatening lives with harmful cures. For instance, drinking fish tank additives, bleach, or cow urine to cure COVID-19 is such a fake information circulating through social media [4]. Moreover, it is also investigated that dissemination of misinformation (known to be false implicitly) “poses a major risk to society” [5].

Though finding an effective treatment for COVID-19 continues to keep its urgency, fighting with fake information on social media has the same or probably more importance level to decrease harmful consequences [6]. Since, most of the social media users are exposed to false COVID-19 information in some way, the research to identify the truth of news becomes vital.

Research approaches to identification of truth of information, i.e. fake news detection, may be divided to three groups: i) language approach using morphological (grammar, syntax) properties of the language to determine false information, ii) state-of-the-art machine learning algorithms are trained with specialized datasets to discover genuineness of news and iii) knowledge-based approaches that combine machine learning with knowledge engineering for the same purpose [7].

English is an extensively studied language in terms of fake news identification and particularly COVID-19 misinformation detection. Though Turkish is a widely spoken language, there are limited works about fake news detection in general. To the best of our knowledge, this is the first fake news identification study specific to COVID-19. This is the first contribution and motivation of this study. Furthermore, Turkish COVID-19 fake news identification requires a benchmarking dataset for research. We generated a novel Turkish COVID-19 fake news benchmarking dataset for the first time in the literature and this is the second contribution of our study. Moreover, recent deep learning language transformers, i.e. Bi-directional Encoder Representations from Transformers (BERT) [8], Embeddings from Language Models (ELMo) [9], are rarely evaluated in Turkish NLP studies. We carried out detailed experiments to develop an advanced COVID-19 fake news identification pipeline with the use of latest deep learning language models. This third contribution is further enriched with the use of Turkish language dedicated neural transformer, BERTurk [10]. We also compared the overall performances of transformer models with frequently used baseline machine learning algorithms and recent deep learning algorithms.

The rest of the paper is structured as follows: Section 2 addresses the latest research related to fake news identification particularly focusing on COVID-19 news. We explain the proposed methodology in Section 3, while presenting language preprocessing tasks, feature filtering methods, conventional machine learning algorithms, various novel deep learning algorithms and advanced neural language transformer models. The experiments and their corresponding results are given in Section 4. The overall research ends with conclusion provided in Section 5.

2. Related work

Automatic discrimination of COVID-19 fake news from trustworthy information is an increasingly vital and a challenging task to solve. Application of advanced machine learning frameworks and assemblage of new benchmarking data may be a solution to COVID-19 fake news identification challenge and other similar public health crises [11]. The recent literature about fake news identification with emphasis on COVID-19 is surveyed in the current section.

Previous search concluded that social media is progressively exploited to manipulate and alter the public opinion [12]. This exploitation on fake news resources, i.e. Reddit or Facebook posts, Twitter tweets etc., are identified with various state-of-the-art machine learning models. Though there are human supported hybrid machine learning models, we basically focus on fully automated identification methods based on recent research. In particular, we classify the relevant COVID-19 fake news identification methods as follows: a) conventional machine learning algorithms such as Naïve Bayes (NB) and Support Vector Machine (SVM), b) Novel deep learning algorithms such as Long Short-Term Memory (LSTM), Bi-directional Long-Short-Term-Memory (Bi-LSTM), Convolutional Neural Networks (CNN), Gated Recurrent Unit (GRU), c) Neural transformers such as BERT, RoBERTa, DistilBERT and BerTURK. We narrowed the literature search results to these categories for the sake of convenience.

a) The conventional algorithms in COVID-19 fake news identification: In [13], SVM model on top of linguistics features such as n-grams and emotional tone is used to discriminate two sentiment classes of COVID-19 Twitter data with F1 score of 95.19%. In a NB based study, the corresponding sentiments of COVID-19 related tweets are determined with accuracy of 91% [14]. Other conventional algorithms, i.e. NB, SVM, Logistics Regression (LR) and Random Forests (RF) are used with resampling strategy to discriminate truth of COVID-19 related data obtained through search engines [15]. Mazzeo et al. obtained accuracy of 94% as their best classification result. In their recent study, Khanday et al. made use of Decision Tree (DT) algorithm with a hybrid feature generation model, i.e. a combination of Frequency-Inverse Document Frequency (TF-IDF), Tweet Length and Bag of Words, to identify state of tweets to be fake or real and they obtained their best classification result as 98.53% [16]. In another conventional algorithm based study [17], Shams et al. made use of various machine learning algorithms such as SVM, NB, RF, LR and Artificial Neural Network (ANN) to detect fake COVID-19 news from Twitter and they obtained accuracy of 93% as their best score with ANN [17].

b) Deep learning neural network architectures in COVID-19 fake news identification: In recent years, numerous deep learning algorithms i.e. LSTM, Bi-LSTM, CNN, GRU and Bi-GRU, are used in COVID-19 fake news detection. In [18], Rustam et al. used Bi-LSTM architecture and Extra Trees Classifiers (ETC) to classify Twitter sentiments related to COVID-19 and they achieved 93% accuracy with ETC. In their Twitter data analysis, Granik et al. made use of CNN and Bi-LSTM neural network architectures with corresponding accuracies of 87% and 86% [19]. CNN and RF algorithms are used to detect real or fake news of COVID-19 from web search and the resultant classification accuracies of both classifiers were obtained as 86.7% [20].

c) Neural transformers in COVID-19 fake news identification: There is an increasing research in COVID-19 NLP applications in terms of usage of neural transformers such as BERT, DistilBERT, RoBERTa. A recent study from literature that makes use of BERT and RoBERTa with 97.1% and 97.9% classification accuracies for “Constraint@AAAI2021-COVID19 Fake News Detection in English” is presented in [21]. The mentioned dataset was also used by Wani et al. to identify fake COVID-19 news with the help of various BERT models and they obtained performance of 98.27% in terms of accuracy [22]. For English COVID-19 benchmarking dataset, Glakova et al. used a special transformer model, i.e. COVID-Twitter-BERT and the resultant accuracy of the transformer was found to be 98.69% [23]. Raha et al. evaluated numerous transformers and baseline algorithms in fake news identification and their best performance was found as 98.2% with RoBERTa transformer [24]. In the study in [25], COVID-19 sentiment analysis for Hindi language was researched and accuracy of BERT model was obtained as 89%. Another study from literature analysis Thai social media to detect COVID-19 fake news

with the use of BERT and GPT models [26]. One last study from literature making use of BERT and ALBERT models as two-stage transformer to identify COVID-19 fake news was studied by Vijjali et al. and this model produced accuracy of 95.6% [11].

Though COVID-19 fake-news study in Turkish language is still limited, there is more recent research in Turkish fake-news identification domain. For instance, traditional algorithms such as SVM, NB, RF, Decision Tree (DT), Extra-Trees Classifier (ETC) were compared in terms of F1-score for fake news obtained from a fact-checking website and they obtained ETC to be the best algorithm [27]. In other recent two studies, Deep Learning and conventional algorithms were compared in terms of accuracy for detection of fake news data [28, 29]. Another study that made use of word-embedding model to identify fake-news is given in the study in [30]. Three recent theses that investigate fake-news identification problem for Turkish were conducted in the studies cited between [31–33], respectively.

Another remarkable point drawn from “Google Scholar” literature search is that the number of publications related to “fake news” increases continuously year by year and this is shown in Figure 1.

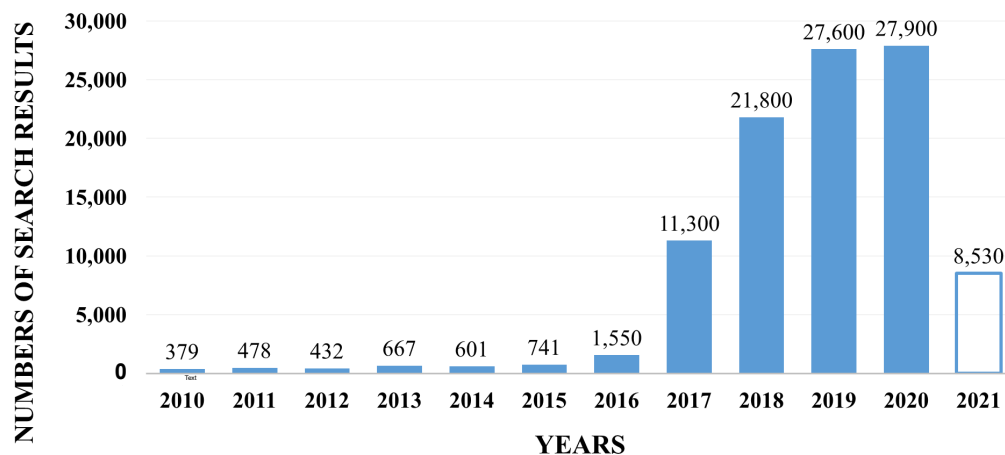


Figure 1. Frequency of research articles for keywords “fake news” from the Google Scholar as of May 30th.

As an overall interpretation, traditional machine learning algorithms combined with advanced preprocessing tasks may give satisfactory results. However, each language processing task is mostly suitable to a particular data and the preprocessing tasks cannot be structured to a generalized pipeline. A shortcoming of traditional machine learning algorithms compared to state-of-the-art (SOTA) algorithms (deep learners and transformers) is that conventional machine learners cannot solve selectivity-invariance problem [34]. Because of this issue, conventional algorithms cannot process data in the original form. In other words, to solve selectivity-invariance problem, they are frequently trained on top of substantial feature selection methods. Since, performance of the conventional algorithms may change depending on the information content of the selected features, many selection methods or their combination should be tested on a trial-and-error mechanism. Contrary of conventional algorithms, deep learners are representational learners that learn features automatically. In other words, deep learners represent original data in higher abstraction levels for decision making. This abstraction mechanism helps to find solution of complex and nonlinear functions. With this nature, transformers are more effective to develop relatively automated language processing pipelines. The data processing superiority of deep learners compared to the conventional algorithms is strictly dependent to the data size used for training. In particular,

the main caveat of deep learners is that as the data-size decreases the performance of the algorithms becomes inadequate. This is where transformers began to rise. Since the transformers were trained with enormous data, even if the data size of the problem is relatively small, the transfer learning technique helps to obtain advanced outcomes for a wide range of language problems, i.e. classification, named entity recognition, topic modeling, word sense disambiguation, text summarization etc.

As it is clear from literature survey, the research and applications in COVID-19 fake news identification is focused on various deep learning architectures and neural language transformers. Conventional algorithms are slightly chosen for baseline purposes. Furthermore, it is also observed that the most of the researched language in this domain is English and other languages constitute only a minor fraction and we should also underline once more that there is no previous research conducted in Turkish COVID-19 fake news domain to the best of our knowledge.

3. Methodology

The methodology followed in this research is presented in this section. The research framework consists of three main tasks: i) generation of dataset from various sources, preprocessing and validation of data ii) investigation of three groups of algorithms and iii) evaluation of results with performance and statistical validation metrics. Hence, this section introduces the workflow mentioned in the three tasks and the whole methodology is summarized in Figure 2.

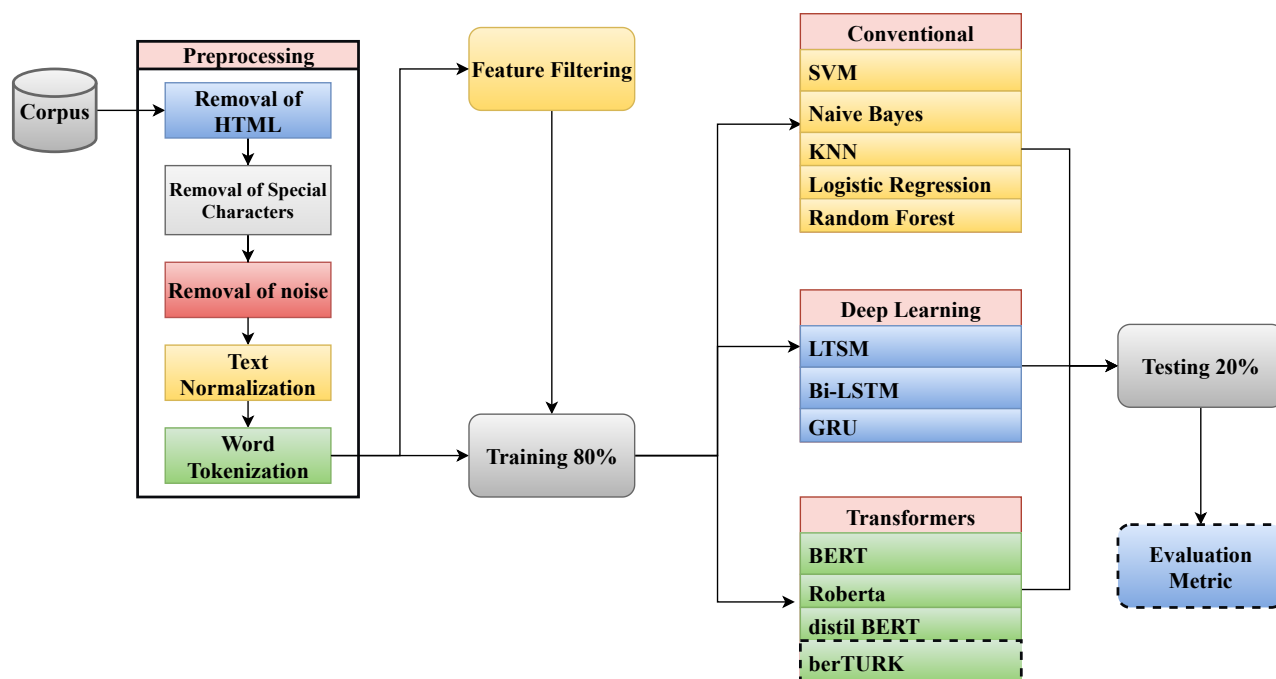


Figure 2. Workflow of COVID-19 fake news identification.

3.1. Dataset collection and preprocessing

In our study, we generated a Turkish dataset from various resources to identify the truth of COVID-19 news. The data, i.e. fake or real news, were obtained from Twitter, Turkish fact-checking web sites: i) www.teyit.org,

ii) www.malumatfurus.org, iii) www.dogrulugune.org and from [35] by translating from English to Turkish. More clearly, we first translated the English COVID-19 data into Turkish with the use of Yandex online machine translation service and then we explored the quality of translation. Since some of the sentences were not relevant for Turkish domain, we manually curated those samples to obtain coherence. The corresponding percentages of the data from the aforementioned resources are given in Table 1 and a few of the sample sentences from the dataset are given in Table 2 respectively.

Table 1. Percentages of the data from the various resources.

Data Source	%
Twitter	12
teyid.org	11
malumatfurus.org	4
dogrulugune.org	3
COVID-19 Dataset [35]	70

Table 2. Sample sentences from generated dataset.

Sample sentences from dataset	English translation of sentences from dataset	Label
Covid-19 aşısını olan İsraili iki saat sonra hayatını kaybetti.	The Israeli, who had the Covid-19 vaccine, died two h later.	Fake
Defin sayımız, 450'lere doğru yürüyor. Normal kasım ayında bizim definimiz, İstanbul'da 202, 190, 180 kişidir. Şu anda 450'ye doğru gidiyor.	Our number of burials is walking towards 450. Our burial in normal November is 202, 190, 180 people in İstanbul. It is heading towards 450 right now.	Real
Türkiye normalleşme döneminde ekonominin tüm göstergelerinde kıyas ülkelerin tamamına göre pozitif ayrılan ülkelere göre biri oldu.	During the normalization period, Turkey became one of the countries that differentiated positively in all economic indicators compared to all benchmarking countries.	Fake
Türkiye toplam hasta sayısına göre ölüm oranı bakımından da Avrupa'daki en iyi ülke konumundadır.	Turkey is also the best country in Europe in terms of mortality rate according to the total number of patients.	Real
Halk sağlığı laboratuvarımızda, Ankara'daki Hıfzıssıha Halk Sağlığı Laboratuvarımızda, merkezimizde aşı üretimi 1998 yılında Sayın Bülent Ecevit hükümeti döneminde tamamen durdurulmuştu.	In our public health laboratory, in our Hıfzıssıha Public Health Laboratory in Ankara, vaccine production in our center was completely stopped in 1998 during the government of Mr. Bülent Ecevit.	Fake

In supervised learning, training of machine learning algorithms is a primary step while developing predictive models. The identification performance of supervised training method strictly depends on the quality of data annotation or labeling. For this aim, first the data is annotated by domain experts. Statistical confidence of data annotation phase requires an Inter-Annotator Agreement (IAA) value to be computed based on the annotation decisions of experts. In this study, we have a two class data and two domain experts to determine labels of dataset. Cohen's Kappa (Kp) is a widely used measure for binary class annotation and two experts [36]. As a reference, While Kp value of 0.61-0.80 indicates significant agreement of annotators, higher values

show perfect agreement. We had collected 2110 COVID-19 samples in total. The two experts both agreed that 1050 samples are true news and 986 are fake news. While selecting the samples into final dataset, a majority vote strategy was applied. In other words, if the two experts agreed for the label of a sample it is retained otherwise it is removed. Following this mechanism 74 samples are removed from entire samples. Based on the annotated samples, the IAA value of Kp was calculated to be 0.942 with the Equation (1). The obtained Kp value indicates a strong-confidence in the annotation phase.

$$Kp = \frac{p_0 - p_e}{1 - p_e}. \quad (1)$$

where in the equation p_0 and p_e are agreement and disagreement percentages of experts.

General statistics of dataset is given in Table 3.

Table 3. Dataset statistics used in analysis.

Category	# of documents	Unique vocabulary	Total vocabulary
Real	1050	8276	26,438
Fake	986	7162	16,492
Totals	2036	15,438	42,930

We preprocessed the extracted data using the following tasks: i) removal of html tags, ii) removal of special characters such as ‘*’, ‘&’, iii) removal of noise, i.e. newlines, white-space, iv) text normalization, i.e. conversion of entire text to lowercase characters, and v) word tokenization. In addition to these steps, a feature selection strategy was applied to observe the corresponding performances of conventional baseline machine learning algorithms.

The conventional algorithms require the input text to be preprocessed before the model is obtained. In this context, the input text should be represented in terms of vector models such as Bag-of-Word (BoW). One of the main drawbacks of BoW models particularly for agglutinative languages is the sparsity and relatively high-dimensional nature. A feature selection strategy, explained in Section 3.2, should be preceded for an efficient model. One another problem of BoW models is their inability to grasp context of the words and hence unable to sense word ambiguity. The concept of the transformers that produce more suitable solutions to these problems, is briefly explained in Section 2 and Section 3.5, respectively.

In the evaluation of the algorithms, we divided dataset into train (80%) and test (20%) splits. While train data is used to identify truth of the news, test split helps to assess performance of the algorithms.

3.2. Feature selection

Text classification performances of conventional machine learning algorithms may change with the number of features that are used in model evaluation. In particular, BoW based NLP models generate high dimensional data that may decrease performance. For this reason, any language modeling may use feature selection techniques, i.e. filters or wrappers, to increase performance of machine learners. Filter methods do not cooperate with learning model and they obtain the best feature subset independently. On the other hand, wrappers select subset of features with learning rules and search algorithms. To further evaluate fake news identification performances of our conventional machine learning algorithms, we made use of Information gain (IG), Gain Ratio (GR), and Correlation Based Feature Selection (CFS) techniques [37, 38].

3.3. Conventional supervised learning methods

As it is observed from literature survey, conventional algorithms are still used as baseline for benchmarking purposes. For this aim, we selected the most frequently used algorithms, i.e. NB, LR, SVM, RF, K-Nearest Neighbor (KNN), to compare the results of advanced algorithms such as deep learners and neural transformers. NB is a probabilistic model that inherits its principle from Bayes theorem with independence assumption between features. NB classifiers obtain the most probable label of a given instance defined by its characteristic vector. NB classifiers differ from Bayes classifiers in that the former learner use features to be independent from the classes [39].

LR is used to solve a wide range of problems and it learns if two variables are linearly related by calculating the strength of the relation. The algorithm is used in many text analysis problems frequently [39].

SVM is successfully used to solve regression and classification problems from various domains. The algorithm is trained to find the optimum separating hyper-plane while separation margin is to be maximized [39]. SVM is effective in nonlinear and high dimensional classification problems.

RF, being an ensemble tree learner, is also effectively used for regression and classification problems. Each tree in RF is trained with subsamples of train set and the prediction is obtained with majority voting for classification [39].

KNN, our last baseline algorithm, is a nonparametric classifier and it forms a neighborhood of instances to be classified with the use of k closest neighbors. The algorithm is also used for regression problems successfully.

3.4. Deep neural network learning models

Deep learning models are frequently used machine learning models having diverse application areas of science. These algorithms are inspired from human nervous system and they contain processing units structured as layers of input, hidden, and output. The processing units in the layers are weighted connected to the units in adjacent layers. The inputs are multiplied with corresponding weights to generate an overall sum. This sum is then feed to an activation function to obtain a resultant decision such as regression or classification [40].

From fake news detection point of view, the most frequently used deep learning algorithms are LSTM, Bi-LSTM, CNN, GRU, and Bi-GRU architectures.

LSTM improves Recurring Neural Network (RNN) in terms of long-term information maintenance and the vanishing gradient issues with the use of gating mechanism. This mechanism decides if the previous state to be retained or not. Derived from LSTM units, Bi-LSTM neural networks operate in bi-directions to oversee past and future information without duplication [41]. LSTM networks particularly Bi-LSTM architectures are frequently used in sequential problems effectively.

CNN architectures are known to be successful in image recognition problems. These networks are recently used in text processing purposes efficiently. In particular, CNN makes use of filters of various sizes that reduces the dimension of sentence matrix. For classification purposes, CNN uses a down-sampling strategy that is known as L2 regularization in tandem with an activation function [42].

GRU, being a modified LSTM model, changes three-gated structure with merging input and forget gates into update gate while the original output gate is used as reset gate. This one-way structure of GRU has issue of losing context information and Bi-GRU remedies this problem with two opposite unidirectional GRUs [43].

3.5. Transformer architectures

The neural transformers introduced effective usage in NLP applications. There are many deep learning based transformer models such as BERT, DistilBERT, RoBERTa, and GPT. In particular, these transformers are frequently used in COVID-19 fake news identification applications.

The mentioned neural models have similar architectures. Therefore we only explain BERT and we refer the users to the literature for the remaining transformer models.

BERT is essentially based on “Masked Language Model” (MLM). In this architecture, MLM masks about 15% percentage of input tokens randomly and the objective of the model is to determine id of the masked vocabulary dependent on the context. The tokens estimated by model are fed into softmax to get the real output word. Furthermore, BERT is pretrained bidirectionally such that it can comprehend left and right context at the same time [8].

In general, BERT is trained in two steps: i) training the model on unlabeled data for diverse pretraining tasks and ii) fine-tuning pretraining parameters of model for the new labeled task such as fake news classification.

BERT is essentially pretrained with English Wikipedia (2500M words) and then its multilingual model generated from various digital sources. The two unsupervised tasks in BERT pretraining are: i) Masking: As it is aforementioned, training phase makes use of input token masking (replacing 15% of token with [MASK] randomly) and then predict the masked token in the output. ii) The second unsupervised training is the next sentence prediction that is helpful in some tasks such as question and answering. The model is trained such that for two sentences A and B; 50% percentage B follows A (labeled as IsNext) and 50% does not follow A (labeled as NotNext) and it is a random sentence from the corpus [8]. The two training phases are shown in Figure 3 and Figure 4, respectively.

Having pretrained the model, BERT architecture can be used for various NLP tasks such as question answering, text summarization, document classification, sentiment analysis, named entity recognition, author identification etc. with a fine-tuning phase. Simply put, task specific inputs and outputs are fed into BERT and then all the parameters are tuned according to the task to be solved. More precisely, rather than training a transformer for each problem, it is computationally more effective to use a pretrained model on the basis of transfer learning.

The relative success of BERT compared to bag-of-words models (BoW) or next generation continuous word embedding models (CBOW) emanate from its bidirectional nature. While BoW models may only be trained with text in right or left order, transformers may process texts bidirectionally for Masked Language Modeling and Next Sentence Prediction tasks at the same time. BoW or advanced CBOW models are inefficient to grasp contextual meaning of words and they are limited to retain the relationships of a specific word with its surroundings. While BoW models do not take into account word orders in a specific text, CBOW models maps one vector for each word that is restricted to grasp contextual meanings of ambiguous words. On the other hand, BERT-like language models process each word depending on other words in a sentence rather than processing them independently. This approach allows BERT to grasp context of the word and it enables framework to discriminate various meaning of polysemous words [8, 9].

BERT has mainly two versions BERT-base and BERT-large and the two versions has 110M and 340M parameters in terms of transformer blocks (L), hidden layers (H), and self-attention heads (A). In this study, we make use of BERT multilingual base model (uncased) in our evaluations and we also measure performance of newly generated Turkish BERT model, i.e. BertTURK [10], for COVID-19 fake news identification. Other

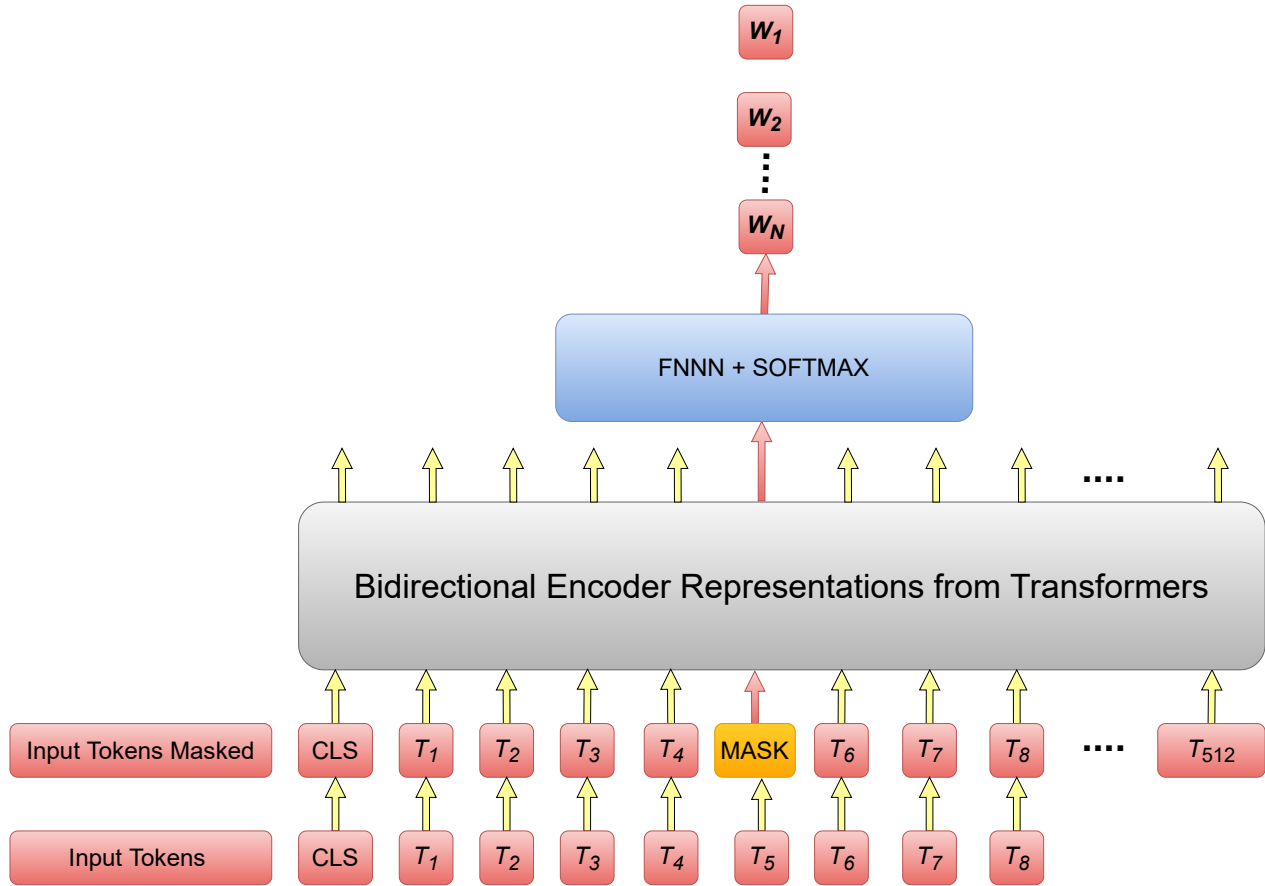


Figure 3. Masking task used in BERT training.

BERT variations utilized in the current study are Distilled BERT (DistilBERT) and Robustly optimized BERT approach (RoBERTa), respectively. While DistilBERT's architecture is the same as BERT, the number of layers is reduced [44]. Furthermore, RoBERTa is another BERT architecture that is optimized with modifying key hyperparameters and training strategies [45]. mBERT, DistilBERT, and mRoBERTa are multilingual and they support Turkish language [46].

3.6. Evaluation and statistical validation metrics

To measure the efficiency of the proposed pipeline, we choose some frequently used evaluation metrics, i.e. Accuracy (Acc), F-1 score (F1), Precision (Pr) and Recall (Rc), from machine learning literature. Furthermore, we statistically validate the outcomes with Matthews's Correlation Coefficient (MCC) and Cohen's Kappa (Kp).

3.6.1. Performance metrics

A classification problem requires metrics to evaluate the outcomes of the algorithms for a detailed comparison. For a binary classification problem, i.e. fake COVID-19 identification, the metrics are derived from Confusion Matrix (CM) given in Figure 5. In the CM, True Positives (TP) and True Negatives (TN) are the correct predictions, while False Negatives (FN) and False Positives (FP) are the incorrect predictions.

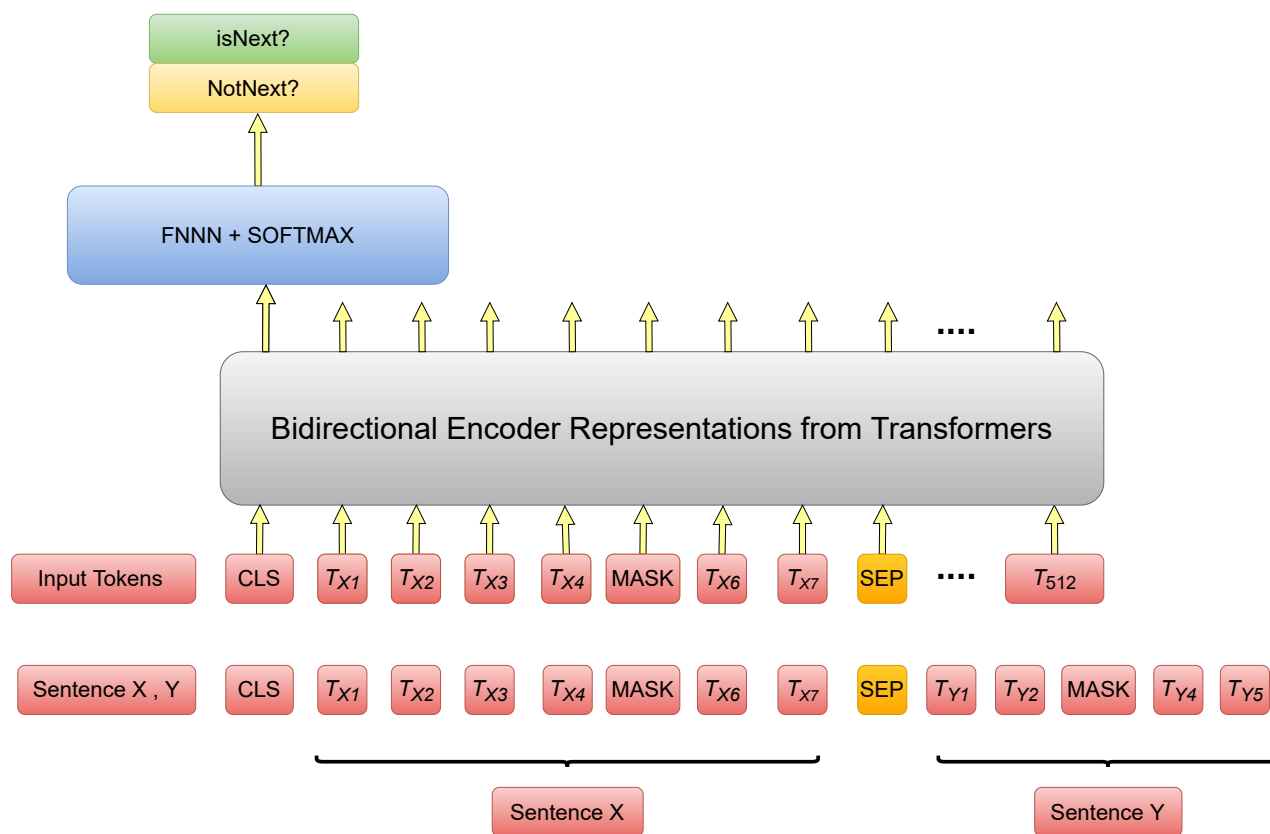


Figure 4. Next sentence prediction in BERT training.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 5. The confusion matrix.

Acc, F1, Pr and Rc metrics in terms of TP, TN, FP and FN are defined as follows [47, 48]: Acc is the correct number of predictions divided by the whole predictions and is given in Equation (2).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Another two important metrics, i.e. Pr and Rc, are defined in Equation (3) and Equation (4), respectively.

$$Pr = \frac{TP}{TP + FP}. \quad (3)$$

$$Rc = \frac{TP}{TP + FN}. \quad (4)$$

while Pr measures the number of positive class predictions that truly is in positive class, Rc calculates the number of positive class predictions out of all positive examples in the dataset. Derived from Pr and Rc another widely used metric, i.e. F1, is given in Equation (5). Though our dataset is almost balanced, we preferred to use micro-averaged F1 score for possible slight variations in outcomes due to the sample sizes.

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = 2 \cdot \frac{Pr \cdot Rc}{Pr + Rc}. \quad (5)$$

Having a balanced benchmarking dataset, we decided to use of Acc and F1 in the evaluation of our experiments.

3.6.2. Statistical validation metrics

Classification outcomes are further assessed with the use of statistical validation metrics, i.e. MCC and Kp. MCC and Kp are also defined in terms of CM and they are given in Equations (6) and (7), respectively.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{((TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN))}}. \quad (6)$$

$$Kp = \frac{p_0 - p_e}{1 - p_e}. \quad (7)$$

where p_0 and p_e are defined in Equations (8) and (9).

$$p_0 = \frac{TP + TN}{TP + TN + FP + FN}. \quad (8)$$

$$p_e = \frac{(TP + FN) \cdot (TP + FP) + (FP + TN) \cdot (FN + TN)}{(TP + TN + FP + FN)^2}. \quad (9)$$

MCC produces values on the [-1,1] range and it is more significant for values approaching to 1. Similarly Kp generates values of [0,1] and the results of the classification is statistically validated for values above 0.8 [37].

4. Experiments and discussions

In this section, we explore the answers of three research questions:

RQ1. How accurate are conventional, deep learning and transformer models to identify COVID-19 fake news?

RQ2. What is the performance of BerTURK compared to multilingual BERT models?

RQ3. Which model achieves overall best COVID-19 fake news identification objective?

4.1. Procedure followed in experiments

Basically, we divide whole experimental procedure in two phases: basic data preprocessing in the evaluation of all algorithms and feature filtering to further assess performance of conventional algorithms. Furthermore, we used the same dataset that was explained in Section 3.1 through all experiments conducted in this study.

While the experiments are carried out, we generated 80% train and 20% data splits for train and testing purposes. Furthermore, while training deep learners and transformers, we additionally split train data to obtain a validation set. Moreover, the parameter sets of algorithms used in the evaluation are given in Table 4.

Table 4. Parameters of machine learning algorithms utilized in evaluation.

Machine learning algorithms	Parameters used in evaluation
Naive Bayes algorithm	Batch size: 100, use kernel estimator: No
Logistic regression classifier	Batch size: 100, ridge: 1.0E-8
Support vector machines	Batch size: 100, cache size: 40.0, cost: 1, degree: 3, loss: 0.1, kernel: linear
Random forest algorithm	Bag size: 100, batch size: 100, number of decimal places: 2, seed: 1
K-nearest neighbor algorithm	Number of neighbors 1, batch size: 100, distance weighting: none, nearest neighbor search method: brute force search, distance measure: Euclidean
LSTM	Vocabulary size:10,000, embedding dimension: 16, maximum length:50, loss: binary crossentropy, optimizer: adam, LSTM node size: 128, drop out: 0.8
Bi-LSTM	Vocabulary size:10,000, embedding dimension: 16, maximum length:50, loss: binary crossentropy, optimizer: adam, Bi-LSTM layers node sizes: 128 and 64, drop out: 0.8
CNN	Vocabulary size:10,000, embedding dimension: 16, maximum length:50, loss: binary crossentropy, optimizer: adam, CNN node size: 128, drop out: 0.5
GRU	Vocabulary size:10,000, embedding dimension: 16, maximum length:50, loss: binary crossentropy, optimizer: adam, GRU node size: 32, drop out: 0.8
Bi-GRU	Vocabulary size:10,000, embedding dimension: 16, maximum length:50, loss: binary crossentropy, optimizer: adam, Bi-GRU node size: 32, drop out: 0.8
BERT	Model name: bert-base-multilingual-cased, maxlen=500, learning rate:2.0E-5 epoch size: 5
RoBERTa	Model name: xlm-roberta-base, maxlen=500, learning rate:2e-5 epoch size: 4
DistilBERT	Model name: distilbert-base-uncased, maxlen=500, learning rate:2.0E-5 epoch size: 5
BerTURK	Model name: bert-base-turkish-cased, maxlen=500, learning rate:2.0E-5 epoch size: 4

4.2. Experimental results

Having generated train-validation-test sets of COVID-19 fake news dataset, we carried out experiments for the parameter sets given in Table 4 and we obtained results in Table 5 and Table 6 in terms of performance metrics. In order to measure confidence of the experiments, we repeated each experiment five times for each algorithm while keeping the parameters constant. We obtained averages of each run and we provide the resultant standard deviations given next to the averages in Table 5 and Table 6.

Table 5. Accuracy and F1 of conventional algorithms.

Conventional Algorithms					
Model	Acc	F1	Model	Acc	F1
NB	.840±.001	.840±.001	SVM-CFS	.830±.011	.830±.011
NB-GR	.886±.003	.887±.003	SVM-IG	.887±.007	.887±.007
NB-CFS	.820±.014	.821±.014	KNN	.619±.019	.605±.019
NB-IG	.840±.001	.840±.001	KNN-GR	.670±.014	.655±.014
LR	.636±.009	.635±.009	KNN-CFS	.813±.005	.812±.005
LR-GR	.673±.012	.673±.012	KNN-IG	.670±.016	.655±.016
LR-CFS	.830±.002	.830±.002	RF	.877±.004	.877±.004
LR-IG	.673±.015	.673±.015	RF-GR	.859±.004	.860±.004
SVM	.869±.007	.870±.007	RF-CFS	.828±.002	.828±.002
SVM-GR	.887±.009	.887±.009	RF-IG	.862±.003	.862±.003

From Table 5, we may conclude that SVM algorithm on IG/GR feature filtered data has the highest performance metrics as 0.887, 0.887 in terms of Acc and F1, respectively. Furthermore, it is observed from Table 5 that NB algorithm with GR filtering has close results to SVM-GR algorithm. Since our benchmarking dataset is balanced, the overall metrics are close to each other in terms of identification/classification quality.

The second and the third group of experiments are devoted to deep learning algorithms and neural transformer models. The obtained results are given in Table 6.

From Table 6, we may observe that deep learning algorithms are in general have better fake news identification ability compared to conventional baseline algorithms. For example, Bi-LSTM, CNN, and Bi-GRU enhance the prediction accuracy and other metrics from 88.7% to 93.1%, 92.4% and 91.7% respectively. Interestingly, transformer based experiment results from Table 6 show great improvement compared to baseline algorithms and deep neural networks. All BERT-like frameworks generate advanced identification results compared to deep neural and baseline algorithms. For example, the lowest prediction accuracy of transformers belongs to DistilBERT with 95.0%. But even this performance indicates a 6.3% and 1.9% performance increase compared to the best results of SVM-GR and Bi-LSTM in terms of accuracy and other metrics. Moreover, the best COVID-19 identification performance belongs to BerTURK with 98.5% accuracy and it is a significant improvement compared to all of the baseline algorithms. Numerically, BerTURK achieves 9.8% increase compared to SVM-GR and 6.7% compared to Bi-LSTM.

We asked in research question 1 to find the algorithm with the best COVID-19 fake news identification performance. From the above results, it is seen that the performance of transformers are better than the remaining algorithms. Specifically, BerTURK has the best accuracy compared to the remaining algorithms. The second research question asks the performance of Turkish language transformer, i.e. BerTURK, compared

Table 6. Accuracy and F1 of deep learning and transformer algorithms.

Deep learning algorithms		
Model	Acc	F1
LSTM	.895±.003	.895±.003
Bi-LSTM	.931±.003	.931±.003
CNN	.924±.002	.924±.002
GRU	.899±.004	.898±.004
Bi-GRU	.917±.015	.916±.015
Transformer models		
Model	Acc	F1
BERT	.973±.002	.972±.002
RoBERTa	.954±.003	.953±.003
DistilBERT	.950±.004	.949±.004
BerTURK	.985±.003	.984±.003

to multilingual BERT versions. It is obvious from the Table 6 that BerTURK improves COVID-19 fake news identification accuracy as much as 1.2% compared to BERT multilingual. We observe that BerTURK with 98.5% accuracy has the best fake news identification ability in overall metrics that is the answer of the third research question. We should also emphasize that since our dataset is balanced, the obtained results are similar in all of the metrics in terms of performance.

Any results obtained from a machine learning problem should also be statistically validated. In this respect, we provide MCC and Kp outcomes of the experiments to validate the certainty of the results in the following section.

4.3. Statistical validation results

In the preceding section, we analyzed the results of 29 experiments in terms of Acc, F1 and we interpreted the results according to these metrics. On the other hand, these results are need to be statistically supported with some validation metrics such as MCC and Kp. In this manner, we calculated MCC and Kp values for all the experiments and we provide the results in Table 7.

As it is observed from Table 7, the high performance algorithms such as BERT models generates significant results in terms of MCC and Kp values. For the sake of convenience, we provide a summarization of top ten results in Figure 6.

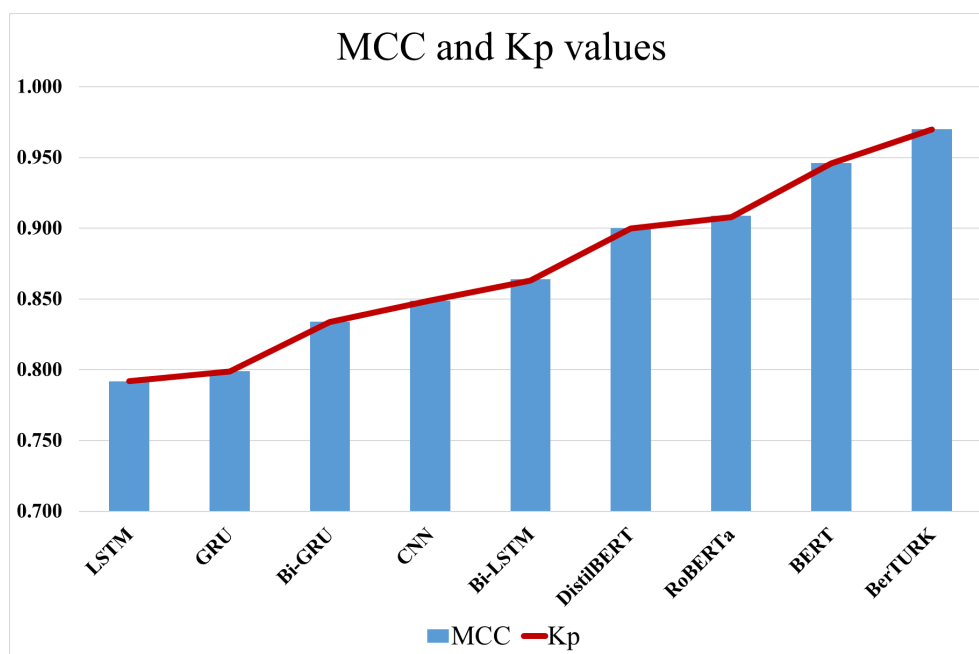
As it can be concluded from Figure 6, the experimental results of transformer models and deep learners all are acceptable in terms of MCC and Kp values. It is again interesting to note that top three algorithms are all have MCC and Kp values above 0.9 that makes their results statistically adequate.

5. Conclusion

In this research, we evaluated performances of 14 machine learning algorithms on a newly collected Turkish COVID-19 fake news dataset. While five of the algorithms are conventional machine learners, the remaining algorithms are deep learners and advanced neural language transformers. The experimental results show that BERT-like models and particularly Turkish language pretrained model (BerTURK) is significantly more accurate

Table 7. MCC and Kp values of all experiments.

Model	MCC	Kp	Model	MCC	Kp	Model	MCC	Kp
NB	.684	.681	SVM-CFS	.665	.661	LSTM	.792	.792
NB-GR	.684	.681	SVM-IG	.776	.774	Bi-LSTM	.864	.863
NB-CFS	.642	.641	KNN	.258	.239	CNN	.849	.849
NB-IG	.684	.681	KNN-GR	.379	.342	GRU	.799	.799
LR	.275	.273	KNN-CFS	.634	.627	Bi-GRU	.834	.834
LR-GR	.347	.347	KNN-IG	.379	.342	BERT	.946	.946
LR-CFS	.664	.661	RF	.755	.754	RoBERTa	.909	.908
LR-IG	.347	.347	RF-GR	.721	.720	DistilBERT	.900	.900
SVM	.741	.740	RF-CFS	.659	.656	BerTURK	.970	.970
SVM-GR	.776	.774	RF-IG	.725	.725			

**Figure 6.** MCC and Kp values of top ten algorithms.

compared to deep neural networks and baseline algorithms. More precisely, BerTURK has achieved a prediction accuracy of 98.5% which outperforms all of the algorithms in terms of identification ability.

Furthermore, deep neural network models such as Bi-LSTM were also observed to be successful in Turkish fake news identification problem compared to baseline algorithms.

Turkish, being morphologically an agglutinate language, is difficult to analyze from NLP problem point of view compared to other widely spoken languages such as English. Any Turkish NLP problem requires preprocessing tasks such as text normalization, feature filtering etc. in order to obtain possible reasonable results on top of conventional machine learners. In spite of the substantial preprocessing tasks, the conventional algorithms may still not be able to generate expected performances. The experimental result also shows that

language transformers may extremely be efficient in solving Turkish computational language problems with minimal preprocessing. This shows a new direction to the researchers.

Another point drawn from the success of transformer architectures is that the transformers are effectively used in various natural language problems with just a minimal fine-tuning strategy. Since the transformers are trained with the use of enormous data, i.e. through digital information, their generalization capacity is superior to the conventional algorithms. On the other hand, since pure health related problems such as drug-disease interactions require high sensitivity, novel medical-domain based transformer architectures are generated and adapted to medical language processing tasks. In order to achieve high-sensitive generalization capability in more specific domains such as medicine, the development of domain based transformers can be a novel research direction.

For future work, we plan to extend our research to design an ensemble based algorithm to further extend COVID-19 fake news detection performance.

References

- [1] Paskin D. Real or Fake News: Who Knows?. *The Journal of Social Media in Society* 2018; 7 (2):252–273.
- [2] Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM et al. The science of fake news. *Science* 2018; 359 (6380):1094–1096.
- [3] Hua J, Shaw R. Corona Virus (COVID-19) Infodemic and Emerging Issues through a Data Lens: The Case of China. *International Journal of Environmental Research and Public Health* 2020; 17(2309):1–11.
- [4] Sear RF, Velasquez N, Leahy R, Restrepo NJ, Oud SE et al. Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning. *IEEE Access* 2020; 8:91886–91893.
- [5] Ciampaglia GL. Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science* 2018; 1 (1):147–153.
- [6] Lampos V, Majumder MS, Yom-Tov E, Edelstein M, Moura S et al. Tracking COVID-19 using online search. *NPJ Digital Medicine* 2021; 4 (1):1–11.
- [7] Beer DB, Matthee M. Approaches to Identify Fake News: A Systematic Literature Review. *Lecture Notes in Networks and Systems*. Springer International Publishing, 2021.
- [8] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* 2018.
- [9] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C et al. Deep contextualized word representation. *arXiv:1802.05365* 2018.
- [10] Schweter S. BERTurk - BERT models for Turkish 2020. doi:10.5281/zenodo.3770924
- [11] Vijjali R, Potluri P, Kumar S, Teki S. Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking. *arXiv:2011.13253v1* 2020.
- [12] Ferrara E. Disinformation and social bot operations in the run up to the 2017 French presidential election. *arXiv:1707.00086* 2017; 22 (8).
- [13] Felber T. Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task. *arXiv:2101.03717* 2021;1–10.
- [14] Samuel J, Ali GGMN, Rahman MM, Esawi E, Samuel Y. COVID-19 public sentiment insights and machine learning for tweets classification. *Information* 2020; 11 (6):1–22.
- [15] Mazzeo V, Rapisarda A, Giuffrida G. Detection of fake news on COVID-19 on Web Search Engines. *arXiv:2103.11804v1* 2021.

- [16] Khanday AMUD, Khan QR, Rabani ST. Identifying propaganda from online social networks during COVID-19 using machine learning techniques. *International Journal of Information Technology* 2021; 13 (1):115–122.
- [17] Shams AB, Apu EH, Rahman A, Raihan MMS, Siddika N et al. Web search engine misinformation notifier extension (Seminext): A machine learning based approach during covid-19 pandemic. *Healthcare* 2021; 9 (2).
- [18] Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A et al. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS ONE* 2021; 16 (2):1–23.
- [19] Granik M, Mesyura V. Fake news detection using naive Bayes classifier. In: 2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, 2017, pp:900–903.
- [20] Choudrie J, Banerjee , Kotecha K, Walambe R, Karende H et al. Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. *Computers in Human Behavior* 2021; 119:1–11.
- [21] Shifath SMS, Khan MF, Islam MS. A transformer based approach for fighting COVID-19 fake news. *arxiv:2101.12027v1* 2021; 1–9.
- [22] Wani A, Joshi I, Khandve S, Wagh V, Joshi R. Evaluating Deep Learning Approaches for Covid-19 Fake News Detection. *arXiv:2101.04012* 2021;153–163.
- [23] Shu K, Wang S, Liu H. Exploiting Tri-Relationship for Fake News Detection Network representation learning View project Feature engineering for outlier detection View project. *arxiv:1712.07709v1* 2017.
- [24] Raha T, Indurthi V, Upadhyaya A, Kataria J, Bommakanti P et al. Identifying COVID-19 Fake News in Social Media. *arXiv:2101.11954v2* 2021.
- [25] Chintalapudi N, Battineni G, Amenta F. Sentimental analysis of COVID-19 tweets using deep learning models. *Infectious Disease Reports* 2021;13 (2):329–339.
- [26] Mookdarsanit P and Mookdarsanit L. The COVID-19 fake news detection in Thai social texts. *Bulletin of Electrical Engineering and Informatics* 2021; 10 (2):988–998.
- [27] Mertoğlu U and Genç B. Automated Fake News Detection in the Age of Digital Libraries. *Information Technology and Libraries* 2020; 39 (4).
- [28] Chauhan NK and Singh K. A Review on Conventional Machine Learning vs Deep Learning. In: 2018 International Conference on Computing, Power and Communication Technologies, 2018.
- [29] Mertoğlu U, Sever H and Genç B. Savunmada Yenilikçi bir Dijital Dönüşüm Alanı: Sahte Haber Tespit Modeli. In:9. Savunma Teknolojileri Kongresi,2018 (in Turkish with an abstract in English).
- [30] Mertoğlu U, Genç B and Sever H. Text-Based Fake News Detection via Machine Learning. In: The International Conference on Artificial Intelligence and Applied Mathematics in Engineering,2020,pp:113-124.
- [31] Taşkın SG.Detecting fake news in Turkish with deep learning algorithms. PhD, Süleyman Demirel University, Isparta, Turkey,2020.
- [32] Özbay FA. Fake News Detection in Online Social Networks Using Swarm Intelligence Based Methods. PhD, Firat University, Elazığ,Turkey,2020.
- [33] Mertoğlu U. Fake News Detection Model For Turkish Language. PhD, Hacettepe University, Ankara, 2020.
- [34] Vasilakos C, Kavroudakis D, and Georganta A. Machine learning classification ensemble of multitemporal Sentinel-2 images: The case of a mixed mediterranean ecosystem. *Remote Sensing* 2020; 12 (12).
- [35] Patwa P, Sharma S,Pykl S,Guptha V, Kumari G et al. Fighting an Infodemic: COVID-19 Fake News Dataset. *arXiv:2011.03327* 2020.
- [36] Ginn R, Pimpalkhute P, Nikfarjam A, Patki MSA, O’Conner K et al. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. In: BIOTXTM’14, 2014.
- [37] Şahin DÖ, Kılıç E. Two new feature selection metrics for text classification. *Automatika* 2019; 60 (2):162–171.

- [38] Borandağ E, Özçift A, Kaygusuz Y. Development of majority vote ensemble feature selection algorithm augmented with rank allocation to enhance Turkish text categorization. *Turkish Journal of Electrical Engineering & Computer Sciences* 2021; 29 (2):514–530.
- [39] Lino FSB, Oliveira MHA, Souza GMF, Rocha LS, Oliveira EL et al. Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis. *Informatics* 2021; 8 (2).
- [40] Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access* 2019; 7:53040–53065.
- [41] Jang B, Kim M, Harerimana G, Kang SU, Kim JW. Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism. *Applied Sciences* 2020; 10 (17).
- [42] Amin MZ, Nadeem N. Convolutional Neural Network: Text Classification Model for Open Domain Question Answering System. *arXiv:1809.02479*, 2018.
- [43] Zhou L, Bian X. Improved text sentiment classification method based on BiGRU-Attention. *Journal of Physics: Conference Series* 2019; 1345 (3).
- [44] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: *EMC2: 5th Edition Co-located with NeurIPS'19*, 2019.
- [45] Liu Y, Ott M, Goyal N, Du J, Joshi M et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [46] Santos DBL, Dutra FGC, Parreiras FS, Brandao WC. Assessing the Effectiveness of Multilingual Transformer-based Text Embeddings for Named Entity Recognition in Portuguese. In: *23rd International Conference on Enterprise Information Systems (ICEIS 2021)*, 1,2021, pp:473-483
- [47] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; 21 (1):1–13.
- [48] Wardhani NWS, Rochayani MY, Iriany A, Sulistyono AD, Lestantyo P. Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data. In: *2019 International Conference on Computer, Control, Informatics and its Applications*, 2019, pp:14–18.