Machine Learning with Applications 4 (2021) 100032

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

A benchmark study of machine learning models for online fake news detection

Junaed Younus Khan [a,1], Md. Tawkat Islam Khondaker [a,1], Sadia Afroz [b], Gias Uddin [c,*], Anindya Iqbal [a]

[a] Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[b] International Computer Science Institute, Berkeley, CA, USA
[c] Department of Electrical and Computer Engineering, University of Calgary, Alberta, Canada

ARTICLE INFO

Keywords:
Fake news
Fake news detection
Benchmark study
Machine learning
Neural network
Deep learning
BERT
Natural language processing

ABSTRACT

The proliferation of fake news and its propagation on social media has become a major concern due to its ability to create devastating impacts. Different machine learning approaches have been suggested to detect fake news. However, most of those focused on a specific type of news (such as political) which leads us to the question of dataset-bias of the models used. In this research, we conducted a benchmark study to assess the performance of different applicable machine learning approaches on three different datasets where we accumulated the largest and most diversified one. We explored a number of advanced pre-trained language models for fake news detection along with the traditional and deep learning ones and compared their performances from different aspects for the first time to the best of our knowledge. We find that BERT and similar pre-trained models perform the best for fake news detection, especially with very small dataset. Hence, these models are significantly better option for languages with limited electronic contents, i.e., training data. We also carried out several analysis based on the models' performance, article's topic, article's length, and discussed different lessons learned from them. We believe that this benchmark study will help the research community to explore further and news sites/blogs to select the most appropriate fake news detection method.

1. Introduction

Fake news can be defined as a type of yellow journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media (Leonhardt & Thompson, 2017). With the growth of online news portals, social-networking sites, and other online media, online fake news has become a major concern nowadays. But people are often unable to spend enough time to cross-check references and be sure of the credibility of news. Hence, considering the scale of the users and contributors to the online media, automated detection of fake news is probably the only way to take remedial measures, and therefore currently receiving huge attention from the research community.

Several research works have been carried out on automated fake news detection using both traditional machine learning and deep learning methods over the years (Dai et al., 2020; Khattar et al., 2019; Rubin et al., 2016; Shu et al., 2019; Tacchini et al., 2017; Wang, 2017; Zhou & Zafarani, 2019). However, most of them focused on detecting news of particular types (such as political). Accordingly, they developed their models and designed features for specific datasets that match their topic of interest. These approaches might suffer from dataset bias and perform poorly on news of another topic. Hence, it is important to study if these are sufficient for different types of news published in online media by evaluating various models on different diverse datasets and comparing their performances. However, the existing comparative studies on fake news detection methods also focused on a specific type of dataset or explored a limited number of models. For example, Wang built a benchmark dataset namely, Liar, and experimented some existing models on it (Wang, 2017). However, the length of this dataset is not sufficient for neural network based advanced models, and some models were found to suffer from overfitting. Gilda explored a few machine learning approaches but did not evaluate any neural network-based model (Gilda, 2017). Recently, Gravanis et al. evaluated a number of machine learning models on different datasets to

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: 1405051.jyk@ugrad.cse.buet.ac.bd (J.Y. Khan), 1405036.mtik@ugrad.cse.buet.ac.bd (Md.T.I. Khondaker), sadia@icsi.berkeley.edu (S. Afroz), gias.uddin@ucalgary.ca (G. Uddin), anindya@cse.buet.ac.bd (A. Iqbal).

[1] The authors contribute equally to this paper.

https://doi.org/10.1016/j.mlwa.2021.100032
Received 7 October 2020; Received in revised form 15 March 2021; Accepted 15 March 2021
Available online 24 March 2021

address the issue of dataset-bias (Gravanis et al., 2019). However, they also did not explore any deep learning based models in their study. Moreover, very few works have been done to explore advanced pre-trained language models (e.g., BERT, ELECTRA, ELMo) for fake news detection (Jwa et al., 2019; Kula et al., 2020) in spite of their state-of-the-art performances in various natural language processing and text classification tasks (Adhikari et al., 2019; González-Carvajal & Garrido-Merchán, 2020; Li et al., 2019; Liu, 2019; Munikar et al., 2019; Peng et al., 2019; Tenney et al., 2019).

Our study fills this gap by evaluating a wide range of machine learning approaches that include both traditional (e.g., SVM, LR, Decision Tree, Naive Bayes, *k*-NN) and deep learning (e.g., CNN, LSTM, Bi-LSTM, C-LSTM, HAN, Conv-HAN) models on three different datasets. We have prepared a new combined dataset containing 80k news of a great variety of topics (e.g., politics, economy, investigation, healthcare, sports, entertainment) collecting from various sources. To the best of our knowledge, this is the largest dataset used for fake news detection study. We also explored a variety of pre-trained models, e.g., BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), ELECTRA (Clark et al., 2020), ELMo (Peters et al., 2018) in our comparative analysis. To the best of our knowledge, no previous study has incorporated such advanced pre-trained models to compare their performance with other machine learning models on fake news detection task. In particular, we answer the following research questions.

RQ1: How accurate are the traditional machine learning vs deep learning models to detect fake news?

We find that deep learning models generally outperform the traditional machine learning models, Among the traditional learning models, Naïve Bayes which achieves 93% accuracy on combined corpus. Among the deep learning models, Bi-LSTM and C-LSTM show great promise with 95% accuracy on combined corpus.

RQ2: Can the advanced pre-trained language models outperform the traditional and deep learning models?

We investigated pre-trained models like BERT, DistilBERT, RoBERTa, ELECTRA, and ELMo. Overall, these models outperform traditional and deep learning ones. For example, the pre-trained RoBERTa shows 96% accuracy on combined corpus, which is more than the traditional and deep learning models. We also find that BERT and similar transformer-based models (BERT, DistilBERT, RoBERTa, ELECTRA) perform better than ELMo.

RQ3: Which model performs best with small training data?

The superior performance of deep learning and pre-trained models we observed in our datasets could be due to large dataset sizes. However, the construction of a large dataset may not always be possible. We, therefore, attempted to understand whether smaller datasets can still be used to train the models without a considerable reduction in accuracy. We see that the pre-trained models can achieve high performance with very small training dataset compared to traditional or deep learning models. For example, RoBERTa achieved over 90% accuracy with only 500 training data used for fine-tuning while traditional and deep learning models fail to achieve even 80% accuracy with such small dataset (see Fig. 3). In contrast, the best performing traditional learning model Naïve Bayes only achieved 65% accuracy with a sample size of 500 training set. Therefore, our finding can be useful for electronic-resource-limited languages where fake news dataset collections are likely to be small in size. In such cases, based on our observations, pre-trained models are the best option to achieve quality performance for these languages. Note that different languages such as Dutch, Italian, Arabic, Bangla, etc. have pre-trained BERT models (Antoun et al., 2020; de Vries et al., 2019; Polignano et al., 2019) that can be fine-tuned with small fake news dataset to develop detection tool.

Replication Package with code and data is shared online at https://github.com/JunaedYounusKhan51/FakeNewsDetection.

Paper Organizations. The rest of this paper is structured as follows. In Section 2, we compare related research works. In Section 3, we describe our study setup by introducing the datasets, the features, and the models we used in our experiments. Section 4 presents the performance of different models on three datasets and answer three research questions. Section 5 compares the performance and analyzes the misclassified cases. We conclude in Section 6.

2. Related work

Related work can broadly be divided into the following categories: (1) Exploratory analysis of the characteristics of fake news, (2) Traditional machine learning based detection, (3) Deep learning based detection, (4) Advanced language model based detection, and (5) Benchmark studies.

2.1. Exploratory analysis of the characteristics of fake news

Several research works have been done over the years on the characteristics of fake news and its' detection. Conroy et al. mentioned three types of fake news: Serious Fabrications, Large-Scale Hoaxes, and Humorous Fakes (Rubin et al., 2015). They have termed fake news as a news article that is intentionally and verifiably false and could mislead readers (Allcott & Gentzkow, 2017). This narrow definition is useful in the sense that it can eliminate the ambiguity between fake news and other related concepts, e.g., hoaxes, and satires.

2.2. Traditional machine learning based detection

Different traditional machine learning based approaches have been proposed for the automatic detection of fake news. In (Shu et al., 2017), the authors proposed to use linguistic-based features such as total words, characters per word, frequencies of large words, frequencies of phrases, i.e., "n-grams" and bag-of-words approaches (Fürnkranz, 1998), parts-of-speech (POS) tagging for fake news detection.

Conroy et al. argued that simple content-related n-grams and part-of-speech (POS) tagging had been proven insufficient for the classification task (Conroy et al., 2015). Rather, they suggested Deep Syntax analysis using Probabilistic Context-Free Grammars (PCFG) following another work by Feng et al. (Feng et al., 2012) to distinguish rule categories (i.e., lexicalized, non-lexicalized, parent nodes, etc.) for deception detection with 85%–91% accuracy. However, Shlok Gilda reported that while bi-gram TF-IDF yielded highly effective models for detecting fake news, the PCFG features had little to add to the models' efficacy (Gilda, 2017).

Many research works also suggested the use of sentiment analysis for deception detection as some correlation might be found between the sentiment of the news article and its type. Ref. (Rubin et al., 2016) proposed expanding the possibilities of word-level analysis by measuring the utility of features like part of speech frequency, and semantic categories such as generalizing terms, positive and negative polarity (sentiment analysis).

Cliche described the detection of sarcasm on twitter using n-grams, words learned from tweets specifically tagged as sarcastic (Cliche, 2014). His work also included the use of sentiment analysis as well as identification of topics (words that are often grouped together in tweets) to improve prediction accuracy.

2.3. Deep learning based detection

Several research works used deep learning models to detect fake news. Wang et al. built a hybrid convolutional neural network model that outperforms other traditional machine learning models (Wang, 2017). Rashkin et al. performed an extensive analysis of linguistic features and showed promising result with LSTM (Rashkin et al., 2017). Singhania et al. proposed a three-level hierarchical attention network, one each for words, sentences, and the headline of a news article (Singhania et al., 2017). Ruchansky et al. created the CSI model where

Table 1
Comparison between our benchmark study and prior benchmark studies.

Theme	Prior benchmark study	Limitations	Our benchmark study
Experiment and result	Bondielli et al. surveyed the different approaches to automatic detection of fake news in the recent literature (Bondielli & Marcelloni, 2019).	They did not run any experiments and did not report any results.	We experimented with all the models and analyzed their performances.
	Dwivedi et al. presented a literature survey on various fake news detection methods (Dwivedi & Wankhade, 2020).		
	Zhang et al. presented an overview of the existing datasets and fake news detection approaches (Zhang & Ghorbani, 2020).		
Dataset length and diversity	Wang experimented with some existing models on their benchmark dataset namely, Liar (Wang, 2017).	They evaluated the models only on one dataset. Moreover, the length of the dataset was not sufficient, and some models were found to suffer from overfitting.	We evaluated all the methods on three different and diverse datasets.
Range of models explored	Gilda explored a few machine learning approaches for fake news detection (Gilda, 2017).		
	Gravanis et al. evaluated a number of machine learning models on different datasets (Gravanis et al., 2019).	They did not evaluate any deep learning based model.	We explored deep learning and advanced pre-trained language models along with traditional ones.
	Oshikawa et al. compared various existing methods for fake news detection on different datasets (Oshikawa et al., 2018).	They did not explore advanced language models such as BERT, ELECTRA, ELMo, etc	

they have captured text, the response of an article, and the source characteristics based on users' behavior (Ruchansky et al., 2017).

Among the recent works, Shu et al. argued that a critical aspect of fake news detection is the explainability of such detection in (Shu et al., 2019). The authors developed a sentence-comment co-attention sub-network to exploit both news contents and user comments. In this way, the authors focused on jointly capturing explainable check-worthy sentences and user comments for fake news detection. In the work (Khattar et al., 2019), the authors developed a multimodal variational auto-encoder by using a bi-modal variational auto-encoder coupled with a binary classifier for the task of fake news detection. The authors claimed that this end-to-end network utilizes the multimodal representations obtained from the bi-modal variational auto-encoder to classify posts as fake or not. Zhou et al. focused on studying the patterns of spreading of fake news in social networks, and the relationships among the spreaders (Zhou & Zafarani, 2019). Hamdi et al. proposed a hybrid approach to detect misinformation in Twitter (Hamdi et al., 2020). The authors extracted user characteristics using node2vec to verify the credibility of the contents.

2.4. Advanced language model based detection

Currently, Advanced pre-trained language models (i.e., BERT, ELECTRA, ELMo) are receiving great attention for several natural language tasks including text classification (Adhikari et al., 2019; González-Carvajal & Garrido-Merchán, 2020; Li et al., 2019; Liu, 2019; Munikar et al., 2019; Peng et al., 2019; Tenney et al., 2019). However, only a few studies have explored them for fake news detection. For example, Jwa et al. detected fake news by analyzing the relationship between the headline and the body text of news (Jwa et al., 2019). The authors claimed that the deep-contextualizing nature of BERT improves F-score by 0.14 over the previous state-of-the-art models. Kula et al. presented a hybrid architecture connecting BERT with RNN to tackle the impact of fake news (Kula et al., 2020). Lee et al. worked on hyperpartisan dataset and leveraged BERT on semi-supervised pseudo-label dataset (Lee et al., 2019).

2.5. Benchmark studies

While most of the existing researches have focused on defining the types of fake news and suggesting different approaches to detect them, very few studies are carried out to compare such approaches independently on different datasets. Among the categories, the benchmark-based studies are the most similar to our study. Table 1 compares our work with the previous benchmark-based studies along three themes: (1) experimental setup and results, (2) dataset length and diversity, and (3) range of models explored. We discuss the related work below.

Wang et al. compared the performance of SVM, LR, Bi-LSTM, and CNN models on their proposed dataset "LIAR" (Wang, 2017). Oshikawa et al. compared various machine learning models (e.g., SVM, CNN, LSTM) for fake news detection on different datasets (Oshikawa et al., 2018). Gravanis et al. compared several traditional machine learning models (i.e., *k*-NN, Decision Tree, Naive Bayes, SVM, AdaBoost, Bagging) for fake news detection on different datasets (Gravanis et al., 2019). Dwivedi et al. presented a literature survey on various fake news detection methods (Dwivedi & Wankhade, 2020). Zhang et al. presented a comprehensive overview of the existing datasets and approaches proposed for fake news detection in previous literature (Zhang & Ghorbani, 2020).

In summary, these few existing comparative studies lack in terms of the range of evaluated models and the diversity of the used datasets. Moreover, a complete exploration of the advanced pre-trained language models for fake news detection and comparison among them and with other models (i.e., traditional and deep learning) were missing

Table 2
Properties of datasets.

Dataset	#Total data	#Fake news	#Real news	Avg. length of news articles (in words)	Topic(s)
LIAR	12791	5657	7134	18	Politics
Fake or real news	6335	3164	3171	765	Politics (2016 USA election)
Combined corpus	79548	38859	40689	644	Politics, economy, investigation, health, sports, entertainment

in previous works. The benchmark study presented in this paper is focused on dealing with the above issues. We extend the state-of-the-art research in fake news detection by offering a comprehensive an in-depth study of 19 models (eight traditional shallow learning models, six traditional deep learning models, and five advanced pre-trained language models).

3. Study setup

In this section, we first introduce the datasets used in our study and discuss how we preprocess those (Section 3.1). Then we discuss different features that we used in our models in Section 3.2. Finally, we discuss the traditional learning, deep learning and pre-trained models that we investigated in our study (Section 3.3). Finally, we discuss the performance metrics we used to evaluate the models and the train and test data settings in Section 3.4.

3.1. Studied datasets

In this comparative study, we make use of three following datasets. Table 2 shows the detailed statistics of them. We describe the datasets below.

3.1.1. Liar

Liar[2] is a publicly available dataset that has been used in (Wang, 2017). It includes 12.8K human-labeled short statements from POLITIFACT.COM. It comprises six labels of truthfulness ratings: pants-fire, false, barely-true, half-true, mostly-true, and true. In our work, we try to differentiate real news from all types of hoax, propaganda, satire, and misleading news. Hence, we mainly focus on classifying news as real and fake. For the binary classification of news, we transform these labels into two labels. Pants-fire, false, barely-true are contemplated as fake and half-true, mostly-true, and true are as true. Our converted dataset contains 56% true and 44% fake statements. This dataset mostly deals with political issues that include statements of democrats and republicans, as well as a significant amount of posts from online social media. The dataset provides some additional meta-data like the subject, speaker, job, state, party, context, history. However, in the real-life scenario, we may not have this meta-data always available. Therefore, we experiment on the texts of the dataset using textual features.

3.1.2. Fake or real news

Fake or real news dataset is developed by George McIntire. The fake news portion of this dataset was collected from Kaggle fake news dataset[3] comprising news of the 2016 USA election cycle. The real news portion was collected from media organizations such as the New York Times, WSJ, Bloomberg, NPR, and the Guardian for the duration of 2015 or 2016. The GitHub repository of the dataset includes around 6.3k news with an equal allocation of fake and real news, and half of the corpus comes from political news.

3.1.3. Combined corpus

Apart from the other two datasets, we have built a combined corpus that contains around 80k news among which 51% are real, and 49% are fake. One important property of this corpus is that it incorporates a wide range of topics including national and international politics, economy, investigation, health-care, sports, entertainment, and others. To demonstrate the topic diversity, we show the inter-topic distances[4] of our combined corpus using LDA-based (Latent Dirichlet Allocation) topic modeling (Blei et al., 2003) in Fig. 1. Based on the empirical analysis of inter-topic distances, we divided the dataset into ten clusters (circles) where each cluster represents a topic. The coordinates of each topic cluster (circle) were measured following the MDS (Multidimensional Scaling) algorithm (Carroll & Arabie, 1998). X-axis (PC1) and Y-axis (PC2) maintained an aspect ratio to 1 to preserve the MDS distances. We used Jensen–Shannon divergence (Lin, 1991) to compute distances between topics. The area of a cluster was calculated by the portion of tokens that respective topic generated compared to the total tokens in the corpus. We named the topic of a cluster based on the most relevant terms representing that cluster. The most relevant terms were determined on the basis of frequency. For example, the most relevant (i.e., most frequent) terms for cluster-7 are 'Trump', 'Clinton', 'Election', 'Campaign', etc (Fig. 1). Hence, the news of this cluster represents the 2016 US election. On the other hand, the most relevant terms for cluster-3 are 'Bank', 'Job', 'Financial', 'Tax', 'Market', etc. Thus, this cluster is related to the Economy. Additionally, overlapping of clusters (e.g., Economy and Politics) indicates shared relevant words (e.g., 'Government', 'People') between them. We have collected news from several sources of the same time domain mostly from 2015 to 2017.[5,6,7] Multiple types of fake news such as hoax, satire, and propaganda have come from The Onion, Borowitz Report, Clickhole, American News, DC Gazette, Natural News, and Activist Report. We have collected the real news from the trusted sources like the New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, New York Post, the Guardian, NPR, Gigaword News, Reuters, Vox, and the Washington Post.

3.1.4. Data preprocessing

Before feeding into the models, raw texts of news required some preprocessing. We first eliminated unnecessary IP and URL addresses from our texts. The next step was to remove stop words. After that, we cleaned our corpus by correcting the spelling of words. We split every text by white-space and remove suffices from words by stemming them. Finally, we rejoined the word tokens by white-space to present our clean text corpus which had been tokenized later for feeding into the models.

3.2. Studied features

We used lexical and sentiment features, n-gram, and Empath generated features for traditional machine learning models, and pre-trained word embedding for deep learning models.

[2] https://www.cs.ucsb.edu/~william/data/liar_dataset.zip.
[3] https://www.kaggle.com/mrisdal/fake-news.
[4] Generated using pyLDAvis: https://pyldavis.readthedocs.io/.
[5] https://homes.cs.washington.edu/~hrashkin/factcheck.html.
[6] https://github.com/suryattheja/Fake-news-detection.
[7] https://www.kaggle.com/snapcrack/all-the-news.

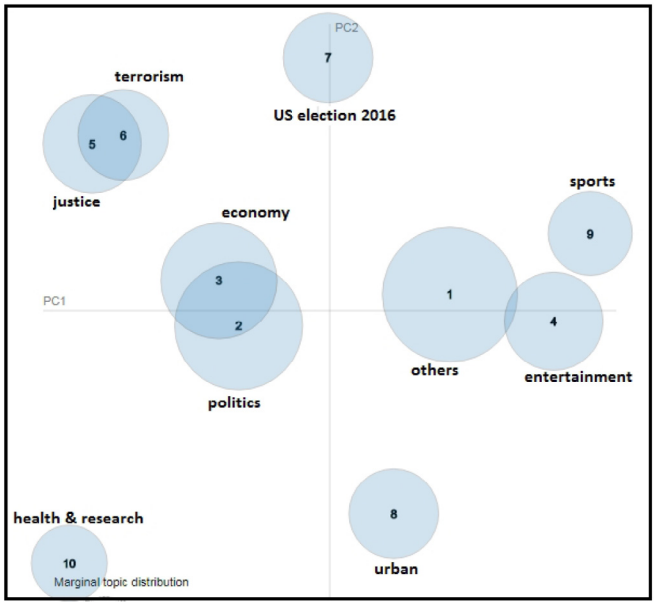

Fig. 1. Inter-topic distance map of Combined Corpus.

3.2.1. Lexical and sentiment features

Several studies have proposed to use lexical and sentiment features for fake news detection (Rashkin et al., 2017; Rubin et al., 2016; Shu et al., 2017). For lexical features, we used word count, average word length, article length, count of numbers, count of parts of speech, and count of exclamation mark. We calculated the sentiment (i.e., positive and negative polarity) of every article and used them as sentiment features.

3.2.2. n-gram Feature

Word-based n-gram was used to represent the context of the document and generate features to classify the document as fake and real (Ahmed et al., 2017; Bourgonje et al., 2017; Granik & Mesyura, 2017; Rashkin et al., 2017; Thorne et al., 2017; Wu & Liu, 2018). We used both uni-gram and bi-gram features in this benchmark and evaluated their effectiveness.

3.2.3. Empath generated features

Empath is a tool that can generate lexical categories from a given text using a small set of seed terms (Fast et al., 2016). Using Empath, we calculated these categories (e.g., violence, crime, pride, sympathy, deception, war) for every news data and used them as features to identify key information in a news article. Since it has been used in literature for understanding deception in review systems (Fast et al., 2016), we feel motivated to investigate their contribution in this context.

3.2.4. Pre-trained word embedding

For neural network models, word embeddings were initialized with 100-dimensional pre-trained embeddings from GloVe (Pennington et al., 2014). GloVe is an unsupervised learning algorithm for obtaining vector representations for words. It was trained on a dataset of one billion tokens (words) with a vocabulary of 400 thousand words.

3.3. Studied models

We experimented various traditional, deep learning and pre-trained language models in this work. Here, we describe all the models that we studied.

3.3.1. Traditional machine learning models

We built our first three models using SVM (Support Vector Machine), LR (Logistic Regression), and Decision Tree with the lexical and sentiment features. Among the four main variants of the SVM kernel, we used the linear one. We also evaluated ensemble learning method like AdaBoost combining 30 decision trees with lexical and sentiment features. Next, we explored the Multinomial Naive Bayes classifier with the n-gram features. We used the Empath generated features with *k*-NN (*k*-Nearest Neighbors) classifier. We use the square-root of the total training data size as *k* as suggested by Lall and Sharma (Lall & Sharma, 1996). Hence, the value of *k* was chosen to be 70, 90, and 250 for Liar, Fake or Real, and Combined Corpus respectively.

3.3.2. Deep learning models

In this study, we have evaluated six deep learning models for fake news detection including CNN, LSTM, Bi-LSTM, C-LSTM, HAN, and Convolutional HAN. The models are described below with their experimental setups.

(1) ***CNN:*** One dimensional convolutional neural network can extract features and classify texts after transforming words in the sentence corpus into vectors (Kim, 2014).The one-dimensional convolutional model was initialized with 100-dimensional pre-trained GloVe embeddings. It contained 128 filters of filter size 3 and a max pooling layer of pool size 2 is selected. A dropout probability of 0.8 was preserved which was expunged for Combined Corpus. The model was compiled with ADAM optimizer with a learning rate of 0.001 to minimize binary cross-entropy loss. A sigmoid activation function was used for the final output layer. A batch size of 64 and 512 was used for training the datasets over 10 epochs.

(2) ***LSTM:*** Our LSTM model was pre-trained with 100-dimensional GloVe embeddings. The output dimension and time steps were set to 300. ADAM optimizer with learning rate 0.001 was applied to minimize binary cross-entropy loss. Sigmoid was the activation function for the final output layer. The model was trained over 10 epochs with batch size 64 and 512.

(3) ***Bi-LSTM:*** Usually, news that is deemed as fake is not fully comprised of false information, rather it is blended with true information. To detect the anomaly in a certain part of the news, we need to examine it both with previous and next events of action. We constructed a Bi-LSTM model to perform this task. Bi-LSTM was initialized with 100-dimensional pre-trained GloVe embeddings. The output dimension of 100 and time steps of 300 was applied. ADAM optimizer with a learning rate of 0.001 was used to minimize binary cross-entropy loss. The training batch size was set to 128 and loss over each epoch was observed with a callback. The learning rate was reduced by a factor of 0.1. We also used an early stop to monitor validation accuracy to check whether the accuracy was deteriorating for 5 epochs. The loss of the binary cross-entropy of the model was minimized by ADAM with a learning rate of 0.0001.

(4) ***C-LSTM:*** The C-LSTM based model contained one convolutional layer and one LSTM layer. We used 128 filters with filter size 3 on top of which a max pooling layer of pool size 2 was set. We fed it to our LSTM architecture with 100 output dimensions and dropout 0.2. Finally, we used sigmoid as the activation function of our output layer.

(5) ***HAN:*** We used a hierarchical attention network consisting of two attention mechanisms for word-level and sentence-level encoding. Before training, we set the maximum number of sentences in a news article as 20 and the maximum number of words in a sentence as 100. In both level encoding, a bidirectional GRU with output dimension 100 was fed to our customized attention layer. We used word encoder as input to our sentence encoder time-distributed layer. We optimized our model with ADAM that learned at a rate of 0.001.

(6) ***Convolutional HAN:*** In order to extract high-level features of the input, we incorporated a one-dimensional convolutional layer before each bidirectional GRU layer in HAN. This layer selected features of each tri-gram from the news article before feeding it to the attention layer.

3.3.3. Advanced language models

Here, we first discuss the advanced language models that we used in this study and then describe their experimental setup.

(1) ***BERT:*** BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model which was designed to learn contextual word representations of unlabeled texts (Devlin et al., 2018). Among the two versions of BERT (i.e., BERT-Base and BERT-Large) proposed originally, we used BERT-Base for this study considering the huge time and memory requirements of the BERT-Large model. The BERT-Base model has 12 layers (transformer blocks) with 12 attention heads and 110 million parameters.

(2) ***RoBERTa:*** RoBERTa (Robustly optimized BERT approach), originally suggested in (Liu et al., 2019), is the second pre-trained model that we experimented. It achieves better performance than original BERT models by using larger mini-batch sizes to train the model for a longer time over more data. It also removes the NSP loss in BERT and trains on longer sequences. Moreover, it dynamically changes the masking pattern applied to the training data.

(3) ***DistilBERT:*** DistilBERT (Sanh et al., 2019) is a smaller, faster, cheaper, and lighter version of original BERT which has 40% fewer parameters than the BERT-Base model. Though original BERT models perform better, DistilBERT is more appropriate for production-level usage due to its' low resource requirements. Considering potential users of non-profit blogs and online media, we think low-resource models have a good appeal. Hence, this is worth investigating.

(4) ***ELECTRA:*** ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) (Clark et al., 2020) is a transformer model for self-supervised language representation learning. This model pre-trained with the use of another (small) masked language model. First, a language model takes an input text and randomly masked the text with generated input token. Then, ELECTRA models are trained to distinguish "real" input tokens vs "fake" input tokens generated by the former language model. At small scale, ELECTRA can achieve strong results even when trained on a single GPU.

(5) ***ELMo:*** ELMo (Embeddings from Language Models) is a contextualized word representation learned from a deep bidirectional language model that is trained on a large text corpus Peters et al. (2018). We used the original pre-trained ELMo model proposed by the authors that has 2 bi-LSTM layers and 93.6 million parameters.

Experimental Setup of Advanced Language Models:

We appended a classification head composed of a single linear layer on the top of the pre-trained advanced language models. The architecture of the classifier head is kept simple to focus on what information can readily be extracted from these pre-trained models. We used the respective pre-trained embeddings of the corresponding models (e.g., BERT embeddings, ELECTRA embeddings, ELMo embeddings) as the input of the classification heads and fine-tuned them for the fake news detection task (Fig. 2). We trained them on all the datasets for 10 epochs with a mini-batch size of 32. We applied early stop to prevent our models from overfitting (Prechelt, 1998b). Validation loss was considered as the metric of the early stopping while delta is set to zero (Prechelt, 1998a). We set the maximum sequence length of the input data to 300. For the Combined Corpus dataset, we configured the gradient accumulation steps as 2 due to the large dataset size. We used AdamW optimizer (Loshchilov & Hutter, 2017) with the learning rate set to 4e-5, ß1 to 0.9, ß2 to 0.999, and epsilon to 1e-8 (Devlin et al., 2018; Sun et al., 2019). Finally, we used binary cross-entropy to calculate the loss (Rosasco et al., 2004). We performed the experiments on NVIDIA Tesla T4 GPU provided by Google Colab.

3.4. Evaluation metrics

We created a standard training and test set for each of the three datasets by splitting it in an 80:20 ratio so that different models can be evaluated on the same ground. For the first two datasets (i.e., Liar, Fake or Real), we did the split randomly as they only contain one type of news. On the other hand, as the Combined Corpus covers a wide variety of topics, we took 80% (20%) data from each topic and include them in train (test) set to maintain a balanced distribution of every topic in training and test data.

We report the performance of each model in terms of accuracy, precision, recall, and F1-score. For precision, recall, and F1-score, we considered the macro-average of both class.

In our experiment, we considered real news as 'positive class', and fake news as 'negative class'. Hence, True Positive (TP) means the news is actually real, and also predicted as real while False Positive (FP) indicates that the news is actually false, but predicted as real. True Negative (TN) and False Negative (FN) imply accordingly. Accuracy is the number of correctly predicted instances out of all instances.

$$Accuracy\ (A) = \frac{TP + TN}{TP + FN + TN + FP} \tag{1}$$

Precision is the ratio between the number of correctly predicted instances and all the predicted instances for a given class. For real and fake classes, we presented this metric as P(R) and P(F) respectively. Hence, the macro-average precision, P will be the average of P(R) and P(F).

$$P(R) = \frac{TP}{TP + FP},\ P(F) = \frac{TN}{TN + FN},\ P = \frac{P(R) + P(F)}{2} \tag{2}$$

Recall represents the ratio of the number of correctly predicted instances and all instances belonging to a given class. For real and fake classes, we presented this metric as $R(R)$ and $R(F)$ respectively. Hence, the macro-average recall, R will be the average of $R(R)$ and $R(F)$.

$$R(R) = \frac{TP}{TP + FN},\ R(F) = \frac{TN}{TN + FP},\ R = \frac{R(R) + R(F)}{2} \tag{3}$$

F1-score is the harmonic mean of the precision and recall.

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{4}$$

4. Study results

In this section, we answer three research questions:

RQ1. How accurate are the traditional and deep learning models to detect fake news in our datasets?

RQ2. Can the advanced pre-trained language models outperform the traditional and deep learning models?

RQ3. Which model performs best with small training data?

Previous studies on fake news detection mainly focused on traditional machine learning models. Therefore, it is important to compare their performance with the deep learning models. We address this concern in RQ1. In particular, the goal of RQ1 is to compare the performance of different traditional machine learning models (e.g., SVM, Naive Bayes, Decision Tree) and deep learning models (e.g., CNN, LSTM, Bi-LSTM) on fake news detection. Considering the great success of pre-trained advanced language models on various text classification tasks, it is important to investigate how these models perform on fake news detection compared to the traditional and deep learning models. The answers to RQ2 will offer insights into whether and how the pre-trained

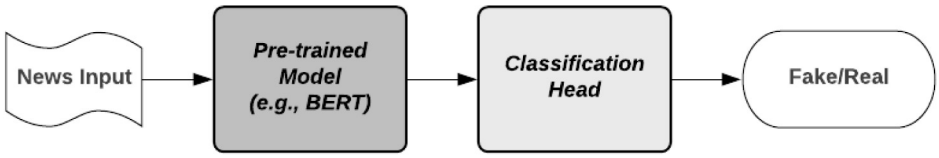

Fig. 2. Fine-tuning of pre-trained language models.

advanced language models are useful to detect fake news. A common issue for any supervised learning problem is the limitation of labeled data. Intuitively, the more performance we can get with less amount of labeled data, the easier it would be to investigate and develop machine learning models to facilitate fake news detection. Therefore, as part of RQ3, we investigate the performance of the models we used in our study on smaller samples of our datasets.

4.1. How accurate are the traditional and deep learning models to detect fake news in our datasets? (RQ1)

In Table 3, we report the performances of various traditional machine learning models in detecting fake news. We observe that among the traditional machine learning models, Naive Bayes with n-gram features performs the best with 93% accuracy on our Combined Corpus. We also find that the addition of sentiment features with lexical features does not improve the performance considerably. For lexical and sentiment features, SVM and LR models perform better than other traditional machine learning models as suggested by most of the prior studies (Chen et al., 2015; Rubin et al., 2016; Tacchini et al., 2017; Wang, 2017; Wu et al., 2017). On the other hand, Empath generated features do not show promising performance for fake news detection, although they had been used earlier for understanding deception in review systems (Fast et al., 2016).

In Table 4, we report the performances of different deep learning models. The baseline CNN model is considered as the best model for Liar in (Wang, 2017), but we find it to be the second-best among all the models. LSTM-based models are most vulnerable to overfitting for this dataset which is reflected by its performance. Although Bi-LSTM is also a victim of overfitting on the Liar dataset as mentioned in (Wang, 2017), we find it to be the third-best neural network-based model according to its performance on the dataset. The models successfully used for text classification like C-LSTM, HAN hardly surmount the overfitting problem for the Liar dataset. Our hybrid Conv-HAN model exhibits the best performance among the neural models for the Liar dataset with 0.59 accuracy and 0.59 F1-score. LSTM-based models show an improvement on the Fake or Real dataset whereas CNN and Conv-HAN continue their impressive performance. LSTM-based models exhibit their best performance on our Combined Corpus where both Bi-LSTM and C-LSTM achieve 0.95 accuracy and 0.95 F1-score. CNN and all hierarchical attention models including Conv-HAN maintain a decent performance on this dataset with more than 0.90 accuracy and F1-score. This result indicates that, although neural network-based models may suffer from overfitting for a small dataset (LIAR), they show high accuracy and F1-score on a moderately large dataset (Combined Corpus).

We find that the traditional machine learning models are generally outperformed by the deep learning models in fake news detection, i.e., the overall accuracy of the traditional models is much lower than the deep learning ones (Tables 3, 4). The difference is more prominent on large dataset, i.e., Combined Corpus which highlights the fact that deep learning models are prone to overfitting on small dataset. However, despite being a traditional model, Naive Bayes (with n-gram) shows great promise in fake news detection which almost reaches the performance of deep learning models and achieves 93% accuracy on Combined Corpus. However, further analysis indicates that the performance of Naive Bayes reaches saturation at some point (2.5K training data) and after that improves very slowly with the increase of sample size, while the performance of the deep learning model, i.e., Bi-LSTM has a greater rate of improvement with the increase of training data (see Fig. 3). So it can be deduced that with enough training samples, deep learning models might be able to outperform Naive Bayes.

Summary of RQ1. How accurate are the traditional vs deep learning models to detect fake news? The deep learning models generally outperform the traditional learning models. The difference of performance between deep learning and traditional models depends on the dataset length. While deep learning models are vulnerable to overfitting on a small dataset, traditional models like Naive Bayes can show impressive performance on this type of dataset. As the dataset length increases, the performance of the deep learning models also improves, and as a result, the deep learning models outperform the traditional models on a large dataset.

4.2. Can the advanced pre-trained language models outperform the traditional and deep learning models? (RQ2)

Table 5 shows the performances of different pre-trained language models on three datasets. While these models incorporate more complex architectures, they do not suffer from overfitting on a smaller dataset as much as the deep learning models do as previously discussed. This is because these models use pre-trained weights in all the layers except the final classification layers. As a result, they do not need a large dataset for fine-tuning their complex architecture. Therefore, all the pre-trained models we evaluated outperform the other traditional ML and deep learning-based models having F1-score no less than 0.62 on the Liar dataset and no less than 0.95 on the Fake or Real News dataset. Given the large dataset (i.e., Combined Corpus), these pre-trained models achieve better performance in the fake news detection task. We observe that among the pre-trained language models, the BERT and transformer-based models (i.e., BERT, RoBERTa, DistilBERT, ELECTRA) are generally better than the other one (i.e., ELMo). For example, DistillBERT (66M parameters), BERT (110M parameters), Electra (110M parameters), RoBERTa (125M parameters), achieve 0.93, 0.95, 0.95, and 0.96 accuracy, respectively on the Combined Corpus dataset while ELMo (93.6M parameters) achieves 0.91. We also notice that the performance of the transformer-based models is proportionate to their number of pre-trained parameters. This relative performance can be justified by their state-of-the-art results on the text classification task (Liu et al., 2019; Sanh et al., 2019).

Summary of RQ2. Can the advanced pre-trained language models outperform the traditional and deep learning models? In our experiment, the pre-trained models perform significantly better than the traditional and deep learning models on all datasets (Tables 3, 4, 5). Since these models are pre-trained to learn contextual text representations on much larger quantities of text corpus and they have produced new state-of-the-art in several text classification tasks (Minaee et al., 2020), their commanding performance over the traditional and deep learning models in the fake news detection task is quite expected.

Table 3
Performance of traditional machine learning models.

Model	Feature	Datasets											
		Liar				*Fake or real news*				*Combined corpus*			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
SVM	Lexical	.56	.56	.56	.48	.67	.67	.67	.67	.71	.78	.71	.72
SVM	Lexical+Sentiment	.56	.57	.56	.48	.66	.66	.66	.66	.71	.77	.71	.72
LR	Lexical+Sentiment	0.56	.56	.56	.51	.67	.67	.67	.67	.76	.79	.76	.77
Decision Tree	Lexical+Sentiment	.51	.51	.51	.51	.65	.65	.65	.65	.67	.71	.69	.7
AdaBoost	Lexical+Sentiment	.56	.56	.56	.54	.72	.72	.72	.72	.73	.74	.73	.74
Naive Bayes	Unigram (TF-IDF)	.60	.60	.60	.57	.82	.82	.82	.82	.91	.91	.91	.91
Naive Bayes	Bigram (TF-IDF)	**.60**	**.59**	**.60**	**.59**	**.86**	**.86**	**.86**	**.86**	**.93**	**.93**	**.93**	**.93**
k-NN	Empath features	.54	.54	.54	.54	.71	.72	.71	.71	.71	.70	.70	.70

Table 4
Performance of deep learning models (using Glove word embedding as feature)

Model	Datasets											
	Liar				*Fake or real news*				*Combined corpus*			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
CNN	.58	.58	.58	.58	.86	.86	.86	.86	.93	.93	.93	.93
LSTM	.54	.29	.54	.38	.76	.78	.76	.76	.93	.94	.93	.93
Bi-LSTM	.58	.58	.58	.57	.85	.86	.85	.85	**.95**	**.95**	**.95**	**.95**
C-LSTM	.54	.29	.54	.38	.86	.87	.86	.86	**.95**	**.95**	**.95**	**.95**
HAN	.57	.57	.57	.56	**.87**	**.87**	**.87**	**.87**	.92	.92	.92	.92
Conv-HAN	**.59**	**.59**	**.59**	**.59**	.86	.86	.86	.86	.92	.92	.92	.92

Table 5
Performance of advanced pre-trained language models.

Model	Datasets											
	Liar				*Fake or real news*				*Combined corpus*			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
BERT	.62	.62	.62	.62	.96	.96	.96	.96	.95	.95	.95	.95
RoBERTa	**.62**	**.63**	**.62**	**.62**	**.98**	**.98**	**.98**	**.98**	**.96**	**.96**	**.96**	**.96**
DistilBERT	.60	.60	.60	.60	.95	.95	.95	.95	.93	.93	.93	.93
ELECTRA	.61	.61	.61	.61	.96	.96	.96	.95	.95	.95	.95	.95
ELMo	.61	.61	.61	.61	.93	.93	.93	.93	.91	.91	.91	.91

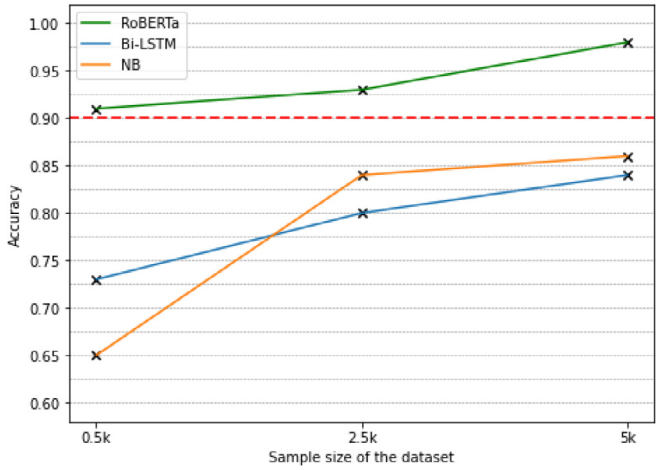

Fig. 3. Comparison of Naive Bayes, Bi-LSTM, and RoBERTa with different training dataset size (from Fake or Real News dataset).

4.3. Which model performs best with small training data? (RQ3)

We find that pre-trained BERT-based models can perform very well with small datasets. We can realize that from their superior performance on small datasets like Liar and Fake or Real News which is significantly better than other models. To further verify this, we take the best model from each of three types, i.e., Naive Bayes with n-gram (traditional), Bi-LSTM (deep learning), RoBERTa (BERT-based) and compare their performances. As their performances differ on the Fake or Real News dataset by very clear margins, we choose this dataset for this analysis. We report their accuracy on small sets of training data (i.e., 500, 2500, and 5000) chosen from Fake or Real News dataset. We show that RoBERTa achieves notably better performance than the other two (Fig. 3). RoBERTa reaches more than 90% accuracy with just 500 training data and continues to improve with the increase of sample size. It hits 98% accuracy with 5000 sample size. On the other hand, both Naive Bayes and Bi-LSTM perform poorly when the size of training dataset is very small, i.e., 500 (Fig. 3). Though their performances improve with the increase of dataset size, they fail to achieve 90% accuracy when the sample size is below 5000.

We further analyze the performance of RoBERTa on smaller datasets (Fig. 4). We find that the model continues to exhibit impressive accuracy (84%) even when the dataset size is 300. This is because pre-trained weights of RoBERTa have already learned the semantic representation from large text corpora. Fine-tuning on the labeled news articles help to learn the model to distinguish between the real and fake news. We observe that the performance of the model starts to drop quickly after the dataset length has been reduced to less than 300. Reducing the data makes it more difficult for the model to differentiate the news articles. Therefore, the performance decreases quickly.

Summary of RQ3. Which model performs best with small training data? Pre-trained models (i.e., RoBERTa) show quality performance even with very small training data in our experiment. We find that RoBERTa achieves over 90% accuracy with a training set of 500 samples only (see Fig. 3).

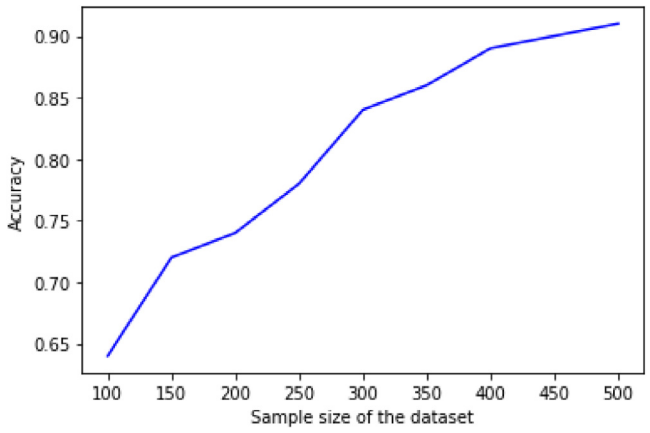

Fig. 4. Comparison of RoBERTa's performance on different training dataset size (from Fake or Real News dataset).

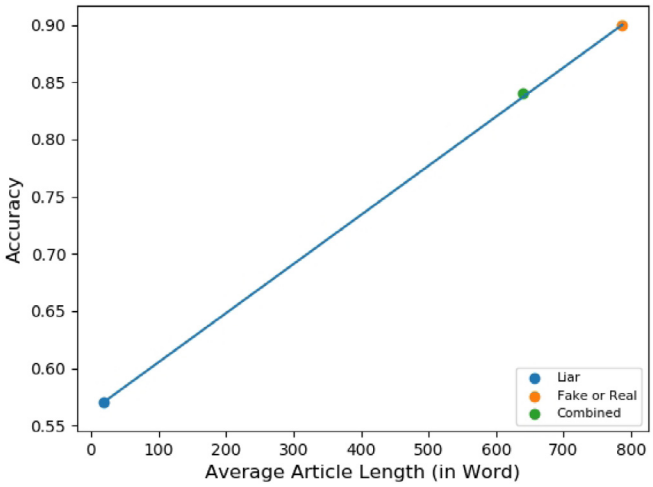

Fig. 5. Relation between models' performance and article length.

5. Discussion

In this section, we compare the performance of the 19 models we studied along several dimensions like features used, resource requirements, etc. (see Section 5.1). We then analyze our models' misclassification, which is discussed in Section 5.2.

5.1. Analysis of performance of different models

In Table 6 we summarize the models we studied in our study based on their accuracy across the three datasets, i.e., Liar, Fake or Real, Combined Corpus. Among the eight types of models we studied under traditional learning approach, Naïve Bayes shows the best accuracy on all the three datasets: combined corpus (0.93), Fake or Real (0.86), and Liar (0.60). Among the six traditional deep learning models we studied, there are three different winners in the three datasets: C-LSTM shows the best performance on the combined corpus (Accuracy = 0.95), HAN shows the best performance (Accuracy = 0.87) on Fake or Real and HAN shows the best performance on the Liar dataset (Accuracy = 0.75). Among the five pre-trained advanced natural language deep learning models we studied, RoBERTa shows the best performance across the three datasets: combined corpus (Accuracy = 0.96), Fake or Real dataset (Accuracy = 0.98), and Liar (0.62). Overall, RoBERTa is the best performing model for two datasets (Combined corpus and Fake or Real) across all the models we studied, while HAN is the best performer for the Liar dataset.

The performance of Naïve Bayes (with n-gram) is only slightly less than the deep learning and pre-trained language models. As such, Naïve Bayes can be a good choice for fake news detection on a sufficiently large dataset with hardware constraint. Naive Bayes (with n-gram) has also been reported to show good performance in spam detection in earlier studies (Hovold, 2005). We find that the performance of Naive Bayes (with n-gram) is almost equivalent to the performances of deep learning models on Combined Corpus (see Table 3). Hence, in the absence of hardware resource requirement of deep learning and advanced pre-trained models (a possible case for non-profit blogs/websites), Naive Bayes with n-gram can be a suitable option with a sufficiently large dataset. Note that the required size of the dataset may vary with its nature, i.e., the number of topics included. However, Naive Bayes fails to achieve considerable accuracy when trained on a minimal sample set (see Fig. 3). Among the diverse features, we studied for the traditional learning models (lexical, sentiment, n-grams), bigram-based models (e.g., Naïve Bayes) show better performance than other features. Overall, the incorporation of sentiment indicators into the models did not improve their performance. For example, for SVM the performance is the same (0.71) for both settings: lexical and lexical + sentiment. Therefore, **Sentiment features are not observed as useful for fake news detection in our study.** The classification of news (as real or fake) has very little to do with the polarity (i.e., sentiment), as fake news can be made up in both directions (positive or negative). While two LSTM-based models (Bi-LSTM, C-LSTM) are the best performer among all the traditional deep learning models, their performance degrades significantly when the dataset sizes are smaller (see RQ3). We observe that LSTM based models show gradual improvement when the dataset length increases from LIAR to Combined Corpus. The more an article contains information, the less these models will be vulnerable to overfitting, and the better they will perform. Hence, neural network-based models may show high performance on a larger dataset over 100k samples (Joulin et al., 2016).

The pre-trained BERT-based models outperform the other models not only on the overall datasets but also on smaller samples of the datasets (see RQ3). We see that the BERT-based model (i.e., RoBERTa) is capable of achieving high accuracy (over 90%) even with a limited sample size of 500 data (see Fig. 3). Hence, these models can be utilized for fake news detection in different languages where a large collection of labeled data is not feasible. Different pre-trained BERT models are already available for different languages, e.g., ALBERTO for Italian (Polignano et al., 2019), AraBERT for Arabic (Antoun et al., 2020), BanglaBERT.[8]

We measured the average training time (per epoch) and GPU usage (during testing) for each BERT-based model on Combined Corpus. We find that the training time needed by DistilBERT is almost half of BERT and RoBERTa, and it requires less GPU for testing (i.e., prediction) as well (see Table 7). Therefore, while DistilBERT shows 0.93 accuracy on the combined corpus which is only slightly behind BERT (0.95) or RoBERTa (0.96), DistilBERT can be useful for production-level usage with hardware constraint and less response time. This is because DistilBERT is developed using the concept of knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015). Hence, it is suitable for production-level usage considering its' high performance and low resource requirement.

5.2. Misclassification analysis

Among the three datasets in our study, the best models (pre-trained language models) show more than 96% accuracy for two datasets (Combined corpus and Fake or Real). For the other dataset (Liar),

[8] https://github.com/sagorbrur/bangla-bert.

Table 6
Summary of all models and performances.

Model type	Model	Rationale for picking	Feature used	Summary of result (Acc.)		
				Liar~	Fake or real	Combined corpus
Traditional machine learning models	SVM	These traditional models are used in different classification tasks including text classification. Different existing studies used them for fake news detection as well.	Lexical	0.56	0.67	0.71
	SVM		Lexical + Sentiment	0.56	0.66	0.71
	LR		Lexical + Sentiment	0.56	0.67	0.76
	Decision Tree		Lexical + Sentiment	0.51	0.65	0.67
	AdaBoost		Lexical + Sentiment	0.56	0.72	0.74
	Naïve Bayes		Unigram	0.60	0.82	0.91
	Naïve Bayes		Bigram	0.60	0.86	0.93
	k-NN		Empath	0.54	0.71	0.71
Deep learning models	CNN	CNN extracts features and classify texts by transforming words into vectors.	GloVe embedding	0.58	0.86	0.93
	LSTM	LSTM remembers information for long sentences.		0.54	0.76	0.93
	Bi-LSTM	Bi-LSTM analyzes a certain part from both previous and next events.		0.58	0.85	0.95
	C-LSTM	Convolutional layer with max-pooling combines the local features into a global vector to help LSTM remembering important information.		0.54	0.86	0.95
	HAN	HAN applies attention mechanism for both word-level and sentence-level representation.		0.75	0.87	0.92
	Conv-HAN	Convolutional layer encodes embedding into feature for word-level and sentence-level attention.		0.59	0.86	0.92
Advanced pre-trained language models	BERT	These language models are~ pre-trained on large text corpus~ and can be fine-tuned for~ text classification.	BERT embeddings	0.62	0.96	0.95
	RoBERTa		RoBERTa embeddings	0.62	0.98	0.96
	DistilBERT		DistilBERT embeddings	0.60	0.95	0.93
	ELECTRA		ELECTRA embeddings	0.61	0.96	0.95
	ELMo		ELMo embeddings	0.61	0.93	0.91

Table 7
Comparison of training time and GPU usage (in testing) for BERT-based models.

Model	#Parameters	Avg. training time per epoch (sec)	GPU used in testing (GB)
DistilBERT	66 M	2175	2.48
BERT	110 M	3149	2.95
RoBERTa	125 M	4020	3.07

the best performing model was HAN with 75% accuracy. Compared to the other two datasets, the Liar dataset has significantly smaller articles (18 words on average) compared to the other two datasets (average 644 words for Combined Corpus and 765 words for Fake or Real news). Indeed, we have observed that when the number of training data is constant, the accuracy of this model is proportional to the average article length of news (see Fig. 5). We confirmed this by analyzing the performance of the Naive Bayes model on 5000 randomly selected records from each of our three datasets. This observation is also consistent with other models. Thus, with the increase of news article length, the models can become more accurate, because those can extract more information to classify the news correctly.

Among the three datasets, two datasets are related to politics (Fake or Real news, Liar), while the other dataset (Combined Corpus) has fake news about diverse topics like health and research, politics, economy, and so on (see Fig. 1 in Section 3). To understand whether the topic of the news has any effect on the classification, we apply topic-based analysis on the fake news articles from the Combined corpus, which our model misclassifies as real. We then map each misclassified case to the ten topics that we found in Fig. 1 of the combined corpus. Overall, quotes are greatly misused to design fake news. We find that the most frequent words in these articles are ‘said’, ‘study’, and ‘research’. The profuse use of the word ‘said’ indicates how fake news sources misconstrue quotes to make these as believable as possible and carry out their own agendas.

Table 8
Topic-wise percentage of false positive news in the Combined Corpus.

Topic	False positive news (%)
Health and research	49.6
Politics	27.6
Miscellaneous	22.8

The topic-wise analysis of misclassification in the combined corpus shows that 49.6% of the false positive news (that are mispredicted as fake in our study) are related to health and research-based topics (Table 8). On the other hand, a tiny portion (27.6%) of the false positive news are related to politics. This high false positive rate of health and research-related news bears evidence that clickbait news on health and research can be produced more convincingly. A slight change in the actual research article will still keep the fake news in the close proximity of the actual article, which makes it difficult to identify them as fake news. In this way, it is quite easy for clickbait news sources to attract people by publishing news claiming the invention of a vaccine for incurable diseases like terminal cancer. Hence, although in recent times the media has focused mostly on combating unauthentic

political news, it should also pay attention to stop the proliferation of false health and research-related news for public safety. We can realize this lesson even better if we think of the impact of fake news during the current COVID-19 pandemic. Corona related fake news has caused serious troubles and confusion among the people. Several fake news such as "Alcohol cures COVID-19", "5G spreads coronavirus", etc have affected people both physically and mentally.[9] Considering the threats associated with it, corona related fake news has been compared to a second pandemic or infodemic.[10]

6. Conclusions

In this study, we present an overall performance analysis of 19 different machine learning approaches on three different datasets. Eight out of the 19 models are traditional learning models, six models are traditional deep learning models, and five models are advanced pre-trained language models like BERT. We find that BERT-based models have achieved better performance than all other models on all datasets. More importantly, we find that pre-trained BERT-based models are robust to the size of the dataset and can perform significantly better on very small sample size. We also find that Naive Bayes with n-gram can attain similar results to neural network-based models on a dataset when the dataset size is sufficient. The performance of LSTM-based models greatly depends on the length of the dataset as well as the information given in a news article. With adequate information provided in a news article, LSTM-based models have a higher probability of overcoming overfitting. The results and findings based on our comparative analysis can facilitate future researches in this direction and also help the organizations (e.g., online news portals and social media) to choose the most suitable model who are interested in detecting fake news. Our future work in this direction will focus on designing models that can detect misinformation and health-related fake news that are prevalent in social media during the COVID-19 pandemic.

CRediT authorship contribution statement

Junaed Younus Khan: Methodology, Wiring - review & editing. **Md. Tawkat Islam Khondaker:** Methodology, Writing - review & editing. **Sadia Afroz:** Writing - review & editing. **Gias Uddin:** Methodology, Writing - review & editing. **Anindya Iqbal:** Methodology, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398.

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using N-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments* (pp. 127–138). Springer.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. arXiv preprint arXiv:2003.00104.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55.

Bourgonje, P., Schneider, J. M., & Rehm, G. (2017). From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism* (pp. 84–89).

Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 535–541).

Carroll, J. D., & Arabie, P. (1998). Multidimensional scaling. In *Measurement, judgment and decision making* (pp. 179–250). Elsevier.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on workshop on multimodal deception detection* (pp. 15–19). ACM.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Cliche, M. (2014). The sarcasm detector. URL: http://www.thesarcasmdetector.com/about.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T annual meeting: information science with impact: Research in and for the community* (p. 82). American Society for Information Science.

Dai, E., Sun, Y., & Wang, S. (2020). Ginger cannot cure cancer: Battling Fake health news with a comprehensive data repository. In *Proceedings of the international AAAI conference on web and social media, Vol. 14* (pp. 853–862).

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). Bertje: A dutch bert model. arXiv preprint arXiv:1912.09582.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dwivedi, S. M., & Wankhade, S. B. (2020). Survey on fake news detection techniques. In *International conference on image processing and capsule networks* (pp. 342–348).

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4647–4657). ACM.

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th annual meeting of the association for computational linguistics: short papers-volume 2* (pp. 171–175). Association for Computational Linguistics.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, *3*(1998), 1–10.

Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. In *Research and development (scored), 2017 IEEE 15th student conference on* (pp. 110–115). IEEE.

González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. In *Electrical and computer engineering (UKRCON), 2017 IEEE first ukraine conference on* (pp. 900–903). IEEE.

Gravanis, G., Vakali, A., Diamantaras, K., & Karadais, P. (2019). Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, *128*, 201–213.

Hamdi, T., Slimi, H., Bounhas, I., & Slimani, Y. (2020). A hybrid approach for fake news detection in Twitter based on user features and graph embedding. In *International conference on distributed computing and internet technology* (pp. 266–280). Springer.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hovold, J. (2005). Naive Bayes spam filtering using word-position-based attributes. In *CEAS* (pp. 41–48).

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences*, *9*(19), 4062.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference* (pp. 2915–2921).

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kula, S., Choraś, M., & Kozik, R. (2020). Application of the BERT-based architecture in fake news detection. In *Conference on complex, intelligent, and software intensive systems* (pp. 239–249). Springer.

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679–693.

Lee, N., Liu, Z., & Fung, P. (2019). Team yeon-zi at semeval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 1052–1056).

Leonhardt, D., & Thompson, S. A. (2017). Trump's lies. *New York Times*, *21*.

Li, X., Bing, L., Zhang, W., & Lam, W. (2019). Exploiting BERT for end-to-end aspect-based sentiment analysis. arXiv preprint arXiv:1910.00883.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145–151.

[9] https://www.bbc.com/news/stories-52731624, Accessed on: Oct 05, 2020.

[10] https://www.nature.com/articles/d41586-020-01409-2, Accessed on: Oct 05, 2020.

Liu, Y. (2019). Fine-tune BERT for extractive summarization. arXiv preprint arXiv:1903.10318.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep learning based text classification: A comprehensive review. arXiv preprint arXiv:2004.03705.

Munikar, M., Shakya, S., & Shrestha, A. (2019). Fine-grained sentiment classification using bert. In *2019 artificial intelligence for transforming business and society (AITB), vol. 1* (pp. 1–5). IEEE.

Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811.00770.

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *CLiC-It*.

Prechelt, L. (1998a). Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, *11*(4), 761–767.

Prechelt, L. (1998b). Early stopping-but when?. In *Neural networks: Tricks of the Trade* (pp. 55–69). Springer.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937).

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M., & Verri, A. (2004). Are loss functions all the same? *Neural Computation*, *16*(5), 1063–1076.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T annual meeting: Information science with impact: Research in and for the community* (p. 83). American Society for Information Science.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7–17).

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 797–806). ACM.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 395–405).

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Singhania, S., Fernandez, N., & Rao, S. (2017). 3HAN: A deep neural network for fake news detection. In *International conference on neural information processing* (pp. 572–581). Springer.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?. In *China National conference on chinese computational linguistics* (pp. 194–206). Springer.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950.

Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., & Vlachos, A. (2017). Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism* (pp. 80–83).

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.

Wu, L., Li, J., Hu, X., & Liu, H. (2017). Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM international conference on data mining* (pp. 99–107). SIAM.

Wu, L., & Liu, H. (2018). Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 637–645). ACM.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), Article 102025.

Zhou, X., & Zafarani, R. (2019). Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter*, *21*(2), 48–60.