Journal of Science and Technology
ISSN: 2456-5660 Volume 7, Issue 09 (November 2022)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2022.v7.i09.pp65-82*

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

Dr.R VVSV PRASAD, Dr. KOPARTHI SURESH

Professor & Head [1], Professor & Principal[2]
Information Technology , Dept. of Computer Science
Swarnandhra College of Engineering and Technology.
Bhimavaram Institute of Engineering And Technology

To Cite this Article
Dr.R VVSV PRASAD, Dr. KOPARTHI SURESH , A benchmark study of machine learning models for online fake newsdetection"
*",Journal of Science and Technology, Vol. 07, Issue 09,-November 2022, pp*65-82
Article Info
Received: 10-09-2022 Revised: 18-09-2022 Accepted: 28-10-2022 Published: 16-11-202

Abstract
The widespread circulation of false information via online platforms is a growing cause for alarm because of the havoc it may wreak. Several machine learning strategies have been proposed for spotting hoaxes. However, the vast majority of them concentrated on a certain category of news (like politics), raising the issue of dataset bias in the used models. Here, we provide the results of a benchmark study that compares three datasets to determine which machine learning technique performs best. To the best of our knowledge, we are the first to investigate and evaluate the performance of many state-of-the-art pre-trained language models for false news detection, alongside the performance of classical and deep learning models. When it comes to detecting false news, we discover that BERT and other comparable pre-trained models perform the best, even when working with a tiny dataset. Because of this, these models are a much superior choice for languages with few electronic contents (i.e., training data). Additionally, we analyzed the models' efficacy, article topics, and article lengths, and shared our findings and insights. We hope that our benchmark study will encourage additional investigation in the field of false news identification and enable news sites and blogs choose the most effective approach.

Keywords*:* Inaccurate reporting Anti-fake news technology Comparison to Industry Standards Learning Machines Connected brains It's a BERT that's been trained using deep learning. Processing of natural language

Introduction

A subset of "yellow journalism" or "pro-paganda," "fake news" describes the dissemination of misleading information via reputable news sources or social media platforms (Leonhardt & Thompson, 2017). The rise of online media such as news websites, social networks, and blogs has contributed to the spread of disinformation. However, few people have the time to double-check information by visiting other sources. With so many individuals accessing and contributing to online material, researchers are putting a lot of effort into finding ways to automatically detect instances of fake news.

Several research, some using traditional machine learning and others using deep learning methods, have been done on automatically identifying bogus news throughout the years (Dai et al., 2020; Khattar et al., 2019; Rubin et al., 2016; Shu et al., 2019; Tacchini et al., 2017; Wang, 2017; Tao et al., 2018).

 Zhou & Zafarani, 2019. However, the vast majority of them sought for just certain types of reports (such as political). They honed down on data sets that would be useful for their studies, and designed features and models appropriately. Due to probable dataset bias, these algorithms may not perform as well when applied to unrelated news articles. Therefore, it is vital to test multiple models on several diverse datasets and compare their performances to ascertain if they are sufficient for various types of news provided in online media. However, prior studies that assessed several techniques for identifying fake news either used a single dataset or only tested a small number of models. Wang is one such example, having developed the Liar benchmark dataset and used it to evaluate current models (Wang, 2017). Some of the neural network models overfitted because the dataset was too little to enable more complex models. Many other types of machine learning

were explored, but no neural network-based model was evaluated (Gilda, 2017). Recently, Gravanis et al. examined the results of many ML models on different data sets.
The data and code described in this research have been verified as reproducible by Code Ocean (see this link: https://codeocean.com/). You may find more information on the Reproducibility Badge Initiative at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals. The writer who is being kept in touch with.
Contact information for the authors is as follows: J.Y. Khan can be contacted at 1405051.jyk@ugrad.cse.buet.ac.bd; Md.T.I. Khondaker can be reached at 1405036.mtik@ugrad.cse.buet.ac.bd; S. Afroz can be reached at sadia@icsi.berkeley.edu; and The best way to get in touch with G. Uddin is via gias.ud (A. Iqbal).

1 Each of the authors has made important contributions to the overall essay.
https://doi.org/10.1016/j.mlwa.2021.100032
In the end, the document was approved on March 15, 2021, after being received on October 7, 2020, and then revised on March 15, 2021.
As of March 24, 2021, you may find this at 2666-8270/ 2021 on the internet. The writer, in his or her own words (s). The publishing firm Elsevier Ltd. is to thank for its availability. As per the provisions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), this work may be read by anyone for free.
datasets that are skewed cause problems (Gravanis et al., 2019). Yet, in their research, deep learning-based models were not explored. Furthermore, despite 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), very few works have been done to investigate advanced pre- We have collected 80,000 news stories on various topics (politics, economics, investigations, healthcare, sports, and entertainment) into a single database. According to our research, this is the largest dataset ever used to study the problem of detecting fake news. Several pre-trained models were also compared: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), Distil- BERT (Sanh et al., 2019), ELECTRA (Clark et al., 2020), and ELMo (Peters et al., 2018). To our knowledge, no previous work has tested the efficacy of different machine learning models on the fake news detection task using such advanced pre-trained models. This section provides extensive answers to the following research questions.
Is there a noticeable difference between the performance of traditional machine learning models and deep learning models when it comes to detecting fake news?

Even though Nave Bayes achieves 93% accuracy on the combined corpus, we find that deep learning models are typically more accurate. Bi-LSTM and C-LSTM are two deep learning models that show promise, with an accuracy of 95% on the combined corpus.
Can cutting-edge, pre-trained language models surpass industry mainstays like as DL and NL?
Pre-trained models such as BERT, DistilBERT, RoBERTa, ELECTRA, and ELMo were investigated. When compared to deep learning and classic models, these ones often perform better. For example, the pre-trained RoBERTa achieves 96% accuracy, which is higher than the accuracy of both traditional and deep learning models on a combined corpus. The transformer-based models BERT, DistilBERT, RoBERTa, and ELECTRA prove superior to ELMo as well.
Third Question: When there is a scarcity of data, which model performs better?
We may have seen better results from deep learning and pre-trained models because of the large datasets we employed. While a large dataset might be useful, it isn't always feasible to create one. So, we investigated the possibility of training the models using less datasets. It is shown that pre-trained models outperform their traditional and deep learning counterparts on much smaller datasets. RoBERTa's accuracy soared to over 90% with only 500 training data used for fine-tuning, whereas the accuracy of both the conventional and deep learning models remained below 80% when given the same size dataset (see Fig. 3). This model is able to learn far more effectively than the best classical learning model, Naive Bayes, which only managed a 65% success rate using a training set of 500 examples. Our findings have relevance for languages where internet availability is more limited, as well as for smaller fake news datasets. According to our findings, using pre-trained models is your best chance for getting reasonable results with these languages. Notably, pre-trained BERT models exist for a number of languages (Antoun et al., 2020; de Vries et al., 2019; Polignano et al., 2019), which might be fine-tuned using a little fake news dataset to create a detection tool.
The replication package's source code and datasets may be found at https://github.com/JunaedYounusKhan51/FakeNewsDetection. This is how the rest of the paper is organized. The second part provides a survey of related research. After that, in the third segment,

present the data, features, and models that served as the foundation for our research. Section 4 presents the findings of the models for three datasets and three research aims. In Section 5, we compare the outcomes and discuss the misunderstandings that occurred. Conclusion is presented in Section 6.
As a second point, the relevant literature
There is a wide range of work that may be categorized as "related," including (1) exploratory analysis of fake news characteristics, (2) conventional machine learning-based detection, (3) deep learning-based detection, (4) advanced language

model-based detection, and (5) benchmark studies.
Critiquing the characteristics that characterize false information
Studies on the characteristics of fake news and how to spot it have been undertaken on several occasions. Conroy and his colleagues propose classifying fabricated stories into three types: severe fabrications, widespread hoaxes, and amusing fakes (Rubin et al., 2015). "Fake news" is what they call articles in the media that are meant to mislead its readers (Allcott & Gentzkow, 2017). This clear definition is useful because it has the potential to clear up any confusion that may exist between fake news and concepts like hoaxes and satire. Detection using traditional methods in machine learning
Several techniques based on traditional machine learning have been published for the automatic detection of fake news. Several linguistic indicators were provided by (Shu et al., 2017) to help spot fake news. These included the number of words in a piece of text, the average number of characters in a word, the frequency of large words and phrases (through "n-grams" and bag-of-words techniques; see also Furnkranz, 1998); and POS tagging.
Document-centric approaches like n-grams and POS tagging, as argued by Conroy et al., are insufficient for this classification task, they say (Conroy et al., 2015). Instead, they suggested that, as Feng et al. before them, Deep Syntax analysis be used to identify deceit with an accuracy of 85-91% by discriminating between different kinds of rules (i.e., lexicalized, non-lexicalized, parent nodes, etc). (Feng et al., 2012). Shlok Gilda found that the PCFG features did not improve the effectiveness of the models for detecting fake news, despite the fact that bi-gram TF-IDF provided incredibly successful models (Gilda, 2017).
Several research have recommended utilizing sentiment analysis for fraud detection since there may be a correlation between the tone of a news story and its category. The authors of the referenced research (Rubin et al., 2016) recommended assessing the worth of variables including part-of-speech frequency and semantic categories such as generalizing terms, positive and negative polarity, and more in order to expand the possibilities of word-level analysis (sentiment analysis).
Cliché proposed using n-grams (words taken from sarcastic tweets) to identify sarcasm on Twitter (Cliche, 2014). To further refine his forecasts, he used sentiment analysis and theme recognition (combinations of words that often appear together in tweets). Forensics propelled by deep learning For example, Wang et al. developed a hybrid convolutional neural network model that vastly outperforms baseline ML techniques at detecting online hoaxes. Positive LSTM findings were found in a comprehensive investigation of linguistic traits done by Rashkin et al (Rashkin et al., 2017). Findings from the research by Singhania et al. suggest that a hierarchical attention network should have three layers: one for individual words, one for phrases, and one for the headline of a news story (Sing- hania et al., 2017). Ruchansky et al. created the CSI model. Assessment of our work against other, more established investigations.
Theme
Prior comparative research
Limitations
Compared to what?

Recent methods for automatically spotting bogus news have been examined by Bondielli et al. (Bondielli & Marcelloni, 2016).
Evidence from an Experiment
2019).
No tests were conducted on this group.
We tried out every single model.
In their review article, Dwivedi et al. provided a variety of techniques for identifying false news (Dwivedi & Wankhade, 2020).
Presenting a survey of current false news detecting methods and data sets, Zhang et al (Zhang & Ghorbani, 2020).
Size and variety of datasets
Wang tried out a few preexisting models on their benchmark dataset, including Liar (Wang, 2017).
Gilda investigated many machine learning strategies for identifying false reports (Gilda, 2017).....and failed to publish their findings.
The models were tested, albeit only on one data set. More importantly, the dataset was too short, and overfitting was seen in a few of the models.
and assessed how they fared.

We tested each strategy on three unique data sets.
Wide variety of models considered
Several machine learning models were tested by Gravanis et al. and compared using a variety of data sets (Gravanis et al., 2019).
Oshikawa et al. examined current techniques for detecting false news across many datasets (Oshikawa et al., 2018).
They didn't put any deep learning-based models through their paces for assessment.

Unlike us, they didn't look at deep learning or pre-trained advanced language models like BERT, ELECTRA, ELMo,

etc.

There is a compilation of text, reader feedback, and article characteristics determined by user behavior (Ruchansky et al., 2017).
To be able to explain how fake news was identified is deemed crucial, according to one of the most recent articles on the issue by Shu et al (Shu et al., 2019). By developing a sentence-comment co-attention sub-network, the authors were able to make use of both news articles and user comments. The writers of this study collaborated to compile user comments and phrases that may be used to illustrate the veracity of fake news. Recently published research (Khattar et al., 2019) describes how to build a multimodal variational auto-encoder for the task of fake news detection by merging a bi-modal variational auto-encoder with a binary classifier. Authors claim that this full-stack network can tell whether a post is fake by analyzing its multimodal representations, which are generated by a bi-modal variational auto-encoder. Zhou et alstudy .'s mainly focused on the dissemination of fake news through social media, specifically on the networks and relationships between people who propagate the misinformation (Zhou & Zafarani, 2019). Twitter is a popular platform for the dissemination of misinformation; Hamdi et al (Hamdi et al., 2020). Extraction of user attributes using node2vec was used to verify this content's accuracy. The use of a cutting-edge language model for detection

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; are just a few of the recent publications that discuss the use of sophisticated pre-trained language models for text categorization and other natural language tasks. Despite a small number of research, these methods are seldom put to use in the fight against fake news. Comparisons between headlines and article bodies have been found by academics like Jwa et al. to be useful in identifying fake news (Jwa et al., 2019). The developers of BERT believe that their model is superior to previous state-of-the-art models in terms of F-score because of its superior contextualization. To cite Kula et al.

combined BERT and RNN architecture to counteract the impact of misinformation (Kula et al., 2020). As part of their work with the hyperpartisan dataset, Lee et al. employed BERT on a semi-supervised pseudo-label data set (Lee et al., 2019). Case studies that will be used as benchmarks
Previous research has mostly focused on classifying false news and developing techniques for identifying it, but there have been very few investigations comparing these approaches independently on different datasets. There was the most conceptual overlap between our two fields and the "benchmark-based research" area. In Table 1, we contrast our findings with those of previous benchmark-based studies in three areas: (1) experimental design and results; (2) dataset length and diversity; and (3) examined models. In the following we elaborate on the research that relates to this topic. By using the LIAR dataset, Wang et al. examined SVM, LR, Bi-LSTM, and CNN models (Wang, 2017). Oshikawa et al. evaluate several machine learning models (including SVM, CNN, and LSTM) for spotting fake news on many datasets (Oshikawa et al., 2018). Gravanis et al. evaluated many different standard machine learning models (i.e. k-NN, Decision Tree, Naive Bayes, SVM, AdaBoost, Bag- ging) on various datasets to see which one was better at spotting fake news (Gravanis et al., 2019). Dwivedi et al. provided a literature study on methods for detecting fake news (Dwivedi & Wankhade, 2020). Zhang et al. performed a comprehensive literature assessment of the existing datasets and approaches for identifying fake news (Zhang & Ghorbani, 2020).
In conclusion, there is a lack of comparative research that adequately covers the range of possible models and datasets. The current state-of-the-art pre-trained language models for fake news detection have not been thoroughly investigated, nor have comparisons been made between these models or with more traditional or deep learning models.

Table 2

Properties of datasets.

Dataset	#Total data articles	#Fake news	#Real news	Avg. length of news (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

Previously, on paper. The purpose of the comparative analysis presented in this work is to answer these questions. In this paper, we contribute to the literature on false news identification by conducting a comprehensive evaluation of 19 current models.

Research Setup

In this section, we first introduce the datasets we used and the methods we used to preprocess them (Section 3.1). The many factors we accounted for in our models are then discussed in Section 3.2. Last but not least, we discuss the various learning methods—from traditional to deep to pre-trained—that we used throughout this study (Section 3.3). Lastly, in Section 3.4, we detail the train and test data setups as well as the performance measures we used to evaluate the models. looked at several datasets

The following are the three data sets that we utilize for evaluation. Table 2 provides an exhaustive count of all of them. Below, we provide specifics on each dataset. The Liar Liar2 dataset is a widely-used tool in a variety of (Wang, 2017). Twelve thousand and eight hundred pithy comments were extracted from POLI- TIFACT.COM and then manually annotated. Pants on fire, false, partly true, half true, mainly true, and true are the six levels of reliability it offers. We make it a point, on a daily basis, to fight against any and all types of false news, propaganda, satire, and misleading reporting. This is why we give special care to identifying genuine news stories from those that are faked. After that, we employ these reformed labels to put news articles into one of two discrete groups. It is generally agreed that everything that is not real is untrue, whereas anything that is real is either entirely t rue or at least somewhat true. When we analyzed the converted data set, we discovered 56% accurate information and 44% erroneous. This dataset focuses mostly on political issues and includes statements from both Democrats and Republicans as well as a large amount of social media posts. The dataset contains meta-data on the subject, speaker, occupation, state, political affiliation, context, and history of the conversation. It's possible that in the real world, we won't always have access to meta-data like this. Thus, we run tests by applying textual features to the texts in the dataset. When reporting news, do media outlets lie or tell the truth?
Whether you want to know if a news story is real or phony, George McIntire makes a data set for you to use. This dataset's fake news component was compiled using information from the Kaggle fake news dataset3. this section was used to analyze stories that were published during the 2016 US election. Publications such as the New York Times, Wall Street Journal, Bloomberg, NPR, and The Guardian from 2015 and 2016 serve as sources for the "real news" section. In all, the dataset on GitHub contains over 6.3k news pieces, with nearly half of the corpus dedicated to political news.

The dataset may be accessed at this link: 2 Here is the link to the dataset: https://www.cs.ucsb.edu/william/data/liar dataset.zip. https://www.kaggle.com/mrisdal/fake-news. Collective textual archive
We have assembled a third dataset of around 80,000 news stories, of which 51% are genuine and 49% are fake. One of the most notable aspects of this corpus is the breadth of topics it covers. By using Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003), we provide the inter-topic distances4 of our combined corpus to demonstrate the plethora of themes. Based on an empirical analysis of their distances, the dataset was split into 10 groups (circles), with each group representing a distinct topic. Through the use of Multidimensional Scaling (MDS), we were able to pinpoint the precise positions of each topical cluster (each circle) (Carroll & Arabie, 1998).
We maintained a 1:1 aspect ratio between PC1 (the X-axis) and PC2 (the Y-axis) to maintain a uniform MDS distance. We used the Jensen-Shannon divergence to evaluate the dissimilarity across groups of study participants (Lin, 1991). The number of tokens in each subject's cluster was calculated as a proportion of the total number of tokens in the corpus. The best sentences were chosen for each cluster and assigned as the topic label. The most crucial terms were determined by analyzing how often they appeared. Cluster 7's most important (most often used) phrases are "Trump," "Clinton," "Election," "Campaign," etc (Fig. 1). In a nutshell, these events constitute the 2016 American presidential campaign. Words like "bank," "job," "finance," "tax," "market," and so on are central to Cluster 3. As a result, the Economy offers a setting in which this community may thrive. In addition, clusters that overlap (such as Economics and Politics) share a great deal of terminology that is relevant to each of their respective fields (like "Government" and "People"). We gathered news from a variety of sources mostly between 2015 and 2017 to cover the same time period. 5 ,6 ,7 A broad range of hoaxes, satires, and pieces of propaganda have been published by outlets including The Onion, Borowitz Report, Clickhole, American News, DC Gazette, Natural News, and Activist Report. The New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, the New York Post, the Guardian, NPR, Gigaword News, Reuters, Vox, and the Washington Post are just some of the reputable sources we combed through to bring you the latest and most accurate news. Preparation and cleaning of data
The models need some preprocessing of the raw texts of news articles. First, we purged the document of any dead links and IP addresses. Then, we cut out all the unnecessary words. After that was complete, we went through our corpus and corrected any instances of misspelling. Every paragraph is separated by white space, and we employ stemming to get rid of all the unnecessary

suffixes. At last, we gave our clean text corpus that had been tokenized and white-space-rejoined for the models to examine. Features taken into account

We used n-grams, Empath, and lexical and sentiment features to train traditional ML models, and we used pre-trained word embeddings to train DL models.[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.

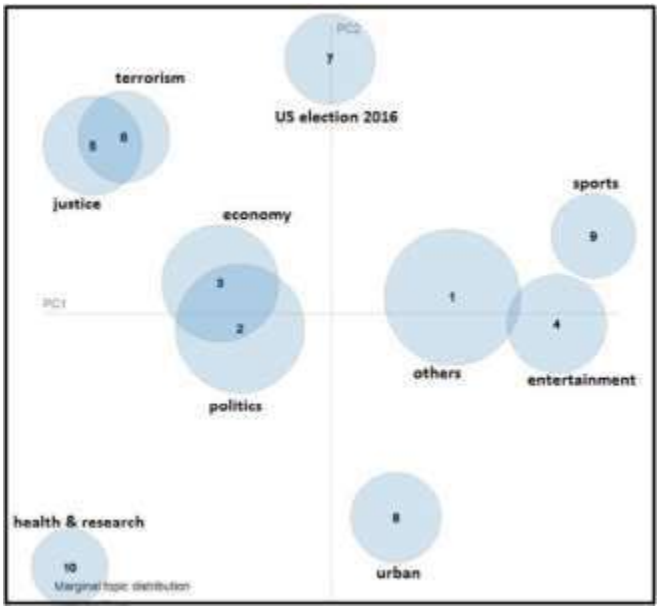

Fig. 1. Inter-topic distance map of Combined Corpus.

Vocabulary and mentality traits
Many studies have suggested using lexical and emotional features as a way to spot fake news (Rashkin et al., 2017; Rubin et al., 2016; Shu et al., 2017). We used the total number of words, average word length, number of words, number of articles, number of different parts of speech, and total number of exclamation marks as lexical features. Each article's positive and negative polarity was calculated, and the numbers were used as sentiment characteristics. Function of N-Grams
We used an n-gram representation of the document's context, which is based on words, to generate features for authenticity detection (Ahmed et al., 2017; Bourgonje et al., 2017; Granik & Mesyura, 2017; Rashkin et al., 2017; Thorne et al., 2017; Wu & Liu, 2018). As a test case, we evaluated the efficacy of uni-gram features against that of bi-gram features. Components developed with sympathy

Empath is a piece of software that, when presented with a text, can extract lexical categories from it using just a small pool of seed words (Fast et al., 2016). Through the use of Empath, we computed a set of categories (such as violence, criminality, pride, pity, deception, and war) for each data point in order to determine which ones were most relevant to the news article in question. Since it has been used in the academic literature to investigate the prevalence of fraud in rating and review systems (Fast et al., 2016), we feel forced to evaluate their usefulness in this field. Using a word embedding that has already been taught
Pre-trained embeddings from GloVe, 100 dimensions in size, were used to seed neural network models with word embeddings (Pennington et al., 2014). GloVe is an unsupervised learning method that may be used to construct word vector representations. It was trained on a dataset of 1 billion tokens, and its vocabulary is 400,000 words (words). seen and analyzed model versions

In this research, we evaluated many linguistic models, including deep learning-based and pre-trained variants. Each of the models we investigated is described in this section. The Most Frequent Machine Learning Techniques
We created three models—a Support Vector Machine (SVM), a Logistic Regression (LR), and a Decision Tree—using the lexical and affective data. Among the four choices, we thought the linear SVM kernel was the best. We also evaluated AdaBoost, an ensemble learning approach that incorporates lexical and sentiment information across 30 decision trees, for its efficacy. Following this, we analyzed the n-gram features using the Multinomial Naive Bayes classifier. The characteristics produced by Empath were sent into the k-NN (k-Nearest Neighbors) classifier. In keeping with Lall and Sharma, we choose k equal to the square root of the total number of training samples (Lall & Sharma, 1996). We used k = 70 for the Liar classification, k = 90 for the Fake or Real classification, and k = 250 for the Combined Corpus classification. Expertise in deep learning for machines

In this study, we evaluated and compared many deep learning models for spotting fake news, including CNN, LSTM, Bi-LSTM, C-LSTM, HAN, and Convolutional HAN. Below, we describe the experimental setups that were utilized to put the models to the test.
To begin, text characteristics and categories may be extracted using a convolutional neural network (CNN) trained on vectors produced from word embeddings (Kim, 2014).
The 1-dimensional convolutional model was seeded with the 100-dimensional GloVe embeddings. It used 128 3-by-3-inch filters and a 2-by-2-inch max pooling layer. In the final combined corpus, a dropout probability of 0.8 was maintained and discarded. The model was built utilizing the ADAM optimizer and a learning rate of 0.001 to minimize the binary cross-entropy loss. A sigmoid function was used to activate the last layer of the network's output. Both 64-element and 512-element batches were used during training, with 10 iterations for each dataset.
The second layer of our LSTM model was pre-trained using 100-dimensional GloVe embeddings. It was decided to keep the time step size at 300 and the output dimension at 300 as well. Using the ADAM optimizer and a learning rate of 0.001, we were able to achieve a minimum binary cross-entropy loss. A sigmoid function was used to activate the last layer of the network's output. The model was trained over 10 iterations using 64- and 512-point batches.
Finally, we have a Bi-Linear Suppressor-Transcriber (Bi-LSTM): There are frequently both true and fraudulent parts to fake news. A news story's outlier may be identified by examining it in the context of similar events, both in the past and in the future. We constructed a Bi-LSTM model for this purpose. The Bi-initial LSTM's seed was a set of pre-trained, 100-dimensional GloVe embeddings. The output was mapped into a 100x100 grid with 300 second intervals. Using the ADAM optimizer and a learning rate of 0.001, we were able to achieve a minimum binary cross-entropy loss. During training, the batch size was held constant at 128, and loss was monitored via callback during the course of each epoch. There was a tenfold decrease in the rate of learning. With an early stop in place, we ran 5 iterations to see whether the validation accuracy dropped. ADAM, with a learning rate of 0.0001, reduced the model's binary cross-entropy loss to a minimum.
Modeling in (4) C-LSTM consisted of a single convolutional layer and a single LSTM layer. Over 128 size-3 filters, we used a maximum pooling layer of size 2. We sent it into our LSTM framework, which has a dropout of 0.2 and 100 output dimensions. Ultimately, a sigmoid function was used to activate the output layer.
A hierarchical attention network (HAN) with two attention processes was used to complete word-level and sentence-level encoding. Training was done using 20-sentence news articles and 100-word sentences. We used a bidirectional GRU with 100-dimensional output space in two-level encoding to feed information into our dedicated attention layer. The time-distributed layer of our sentence encoder relied on the output of the word encoder. Our model was fine-tuned with the use of ADAM's 0.001 epoch learning rate.
(1) Convolutional HAN: For each bidirectional GRU layer in HAN, we added a one-dimensional convolutional layer to extract high-level features of the input. The attention layer received just the tri-grams of the news report that included information that was relevant to the topic at hand. Language-recognizing robots
We introduce the cutting-edge language models and testing infrastructure that underpinned this effort.

A pre-trained model dubbed BERT (Bidirectional Encoder Representations from Trans- formers) has been designed to learn contextual word representations of unlabeled texts (Devlin et al., 2018). Because of its substantially reduced time and memory demands, we choose to use BERT-Base for our analysis, one of two variants proposed in the original BERT proposal. The BERT-Base model has 110 million parameters distributed over 12 levels (transformer blocks).
RoBERTa (Robustly optimized BERT technique), which was initially presented in, was the second pre-trained model we evaluated (Liu et al., 2019). By training the model for longer on more data and with larger mini-batch sizes, it is possible to improve performance relative to baseline BERT models. Moreover, longer sequences are employed in training, and the BERT NSP loss is removed. And when it's being utilized, the masking pattern shifts and changes dynamically.

Due to the massive amount of the data, we will assume that there are only two phases of gradient buildup. Specifically, we used AdamW optimizer (Loshchilov & Hutter, 2017) with the following settings: learning rate = 4e-5, ß1 = 0.9, ß2 = 0.999, and epsilon = 1e-8 to determine the best parameters to use (Devlin et al., 2018; Sun et al., 2019). The last phase in the loss computation was performed using binary cross entropy (Rosasco et al., 2004). The research was conducted on the NVIDIA Tesla T4 GPU housed at Google Colab. Scales and criteria for evaluation
We created a uniform training and test set by splitting each of the three datasets 80:20 so that models could be compared fairly. We arbitrarily split the first two datasets into "Liar," "Fake," and "Real" categories since each contains only one kind of news article. However, given the variety of topics covered by the Combined Corpus, we selected 20% (20%) of the total number of occurrences of that subject for the train set and 100% (100%) for the test set to ensure that we included a representative sample of data from each subject area in both stages.
The performance of the models is reported in a number of different ways, including their accuracy, precision, recall, and F1-score. The F1-score was determined by taking the average performance on the accuracy and recall measures worldwide for both classes.
We classified news as either positive (authentic) or negative (fake) (false). Accordingly, Real Positive (TP) denotes that the news is true and was properly expected as true, while Untrue Positive (FP) denotes that the news is untrue but was accurately predicted as true. Suits True Negative (TN) and False Negative (FN) inferences (FN). Accuracy refers to the rate at which a forecast is correct.

used on training data, or used on.

DistilBERT (Sanh et al., 2019) is a portable and quick distillation instrument.

Accuracy (A) = T P + T N \sT P F N T N F P

(1)

The BERT-Lite model, which has fewer parameters (by 40%) than the BERT-Base model, is both more lightweight and less expensive. Despite the original BERT models' higher performance, DistilBERT is more suited for usage in a production context because to its lower resource requirements. Factoring in the audience of philanthropic blogs and other internet media,

When evaluating a classifier's performance, it is important to look at how many instances it correctly predicts in comparison to how many instances it predicts in total. For real classes, we referred to this metric as P(R), whereas for fake classes, we used P(F). So, P, the average precision, is determined by averaging P(R) and P. (F).

We find ourselves drawn to models that use less resources. This is because

T P

T N

P (R) + P (F)

deserves additional investigation.

P (R) = T P + , P (F) = T N + F N , P = 2 (2) (2) (2)

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) (Clark et al., 2020) is a transformer model for self-supervised language representation learning. A small, standalone masked language model was used to pre-train this model. The first step in training a language model is feeding it some data.

The recall statistic tracks the proportion of correctly predicted instances to the total number of examples in a given class. This metric was represented by the symbols R(R) and R(F) for the correct and incorrect answers, respectively. So, the overall average recall, R, is just the mean of the two recall values, R(R) and R. (F).

text, and then arbitrarily overwrote it with a token based on the input.

(R) = T P , R(F) = T N , R = R(R) + R(F)

(3)

The ELECTRA models are trained to identify "real" information.

T P + F N

T N + F P 2

to input tokens and the "fake" tokens generated by the previous language model. Small-scale training of ELECTRA on a single GPU produces reasonable results.
As for the third example, ELMo (Embeddings from Language Models) by Peters et al. is a deep bidirectional language model trained on a large text corpus that generates a contextualized word representation (2018). To achieve our goals, we used a classic, pre-trained ELMo model from the first generation.

The F1-score is the ideal compromise between precision and recall.
F 1 = 2 · P · R \sP R \s3.
Results and Conclusions
The following are the three research topics we addressed, along with our findings:

(4)

employing their proposed model, which consists of 2 bi-LSTM layers and 93.6 million parameters.
Experimental Setting for Novel Language Models:
We stacked a single-layer linear classification head on top of the state-of-the-art language models that had already been trained. The classifier's neural network has been simplified so it can focus on information that is readily available from pre-trained models. We used the relevant models' pre-trained embeddings as the input of the classification heads and tweaked them for the fake news detection task (Fig. 2). Each dataset was trained using all 32 mini-batch sizes for a total of 10 iterations. Having a hard stop in place at an inappropriate time prevents our models from being too restrictive (Prechelt, 1998b). The validation loss served as the early termination threshold, whereas delta was held constant at 0. (Prechelt, 1998a) There was a cap of 300 items placed on the length of the input sequences. Our goal in creating the Combined Corpus was to

Initial Question: How accurate are our regular and deep learning models in spotting instances of fake news in our datasets? Can the cutting-edge, pre-trained language models outperform the time-tested, ML-based approaches?
RQ3. Which model performs best when there is a scarcity of information?
Traditional machine learning methods have been the focus of much prior research on the detection of fake news. That's why it's so important to evaluate their findings in relation to those of deep learning models. This is addressed in RQ1. To be more specific, RQ1 aims to compare and contrast the performance of multiple deep learning models (such as CNN, LSTM, Bi-LSTM) with that of more traditional machine learning models (such as SVM, Naive Bayes, Decision Tree) when it comes to identifying fake news. It is important to investigate how well pre-trained advanced language models perform in comparison to the tried-and-true methods of detecting false news and the more recent methods of deep learning because of the models' widespread use and apparent efficacy for a wide range of text classification tasks. Further investigation into whether and how the pre-trained

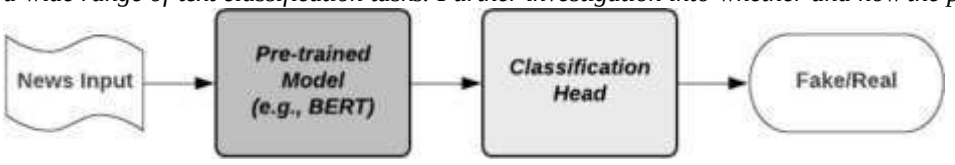

Fig. 2. Fine-tuning of pre-trained language models.

Advanced language models may aid in the detection of fake news. For every supervised learning activity, the lack of labeled data presents an insurmountable obstacle. Improved performance with less labeled data would be helpful in exploring and developing machine learning models to assist with fake news detection. Therefore, as part of RQ3, we conduct tests on subsets of our datasets to see how well our models perform. Can we compare how well traditional and deep learning algorithms spot fake news in our dat a? (RQ1)
We compile the findings of numerous traditional ML models' abilities to detect fake news in Table 3. Naive Bayes with n-gram features outperforms other conventional machine learning models, achieving 93% accuracy on our Combined Cor- pus, to the best of our knowledge. Similarly, we demonstrate that including lexical data along with sentiment traits does not substantially improve performance. For lexical and sentiment features, the present research overwhelmingly favors SVM and LR models over other traditional machine learning methods (Chen et al., 2015; Rubin et al., 2016; Tacchini et al., 2017; Wang, 2017; Wu et al., 2017). However, Empath generated traits do not yield promising outcomes when used to identify fraudulent news, despite their history of use in detecting dishonesty in review systems (Fast et al., 2016).

Table 4 shows a summary of our deep learning experiments' outcomes. Despite the fact that the baseline CNN model is the one recommended by (Wang, 2017) for identifying liars, we rate it lower. Overfitting is a problem in general, but it becomes much more pronounced in LSTM-based models, which is shown in their performance on this dataset. We rank Bi-LSTM as the third-best neural network-based model on the Liar dataset, despite the fact that it has the same overfitting concerns that those highlighted in (Wang, 2017). For all their success with text classification, models like C-LSTM and HAN have trouble with the Liar dataset's notorious overfitting problem. Our hybrid Conv-HAN model has the highest accuracy and F1-score of any neural network model on the Liar dataset (0.59 and 0.60, respectively). While LSTM-based models do better on Fake or Real, CNN and Conv-HAN still provide impressive results. LSTM-based models excel on our Combined Corpus, with Bi-LSTM and C-LSTM both achieving 0.95 accuracy and 0.95 F1-score. Accuracy and F1-scores of 0.90 or higher are maintained across the board by CNN and other hierarchical attention models, notably Conv-HAN, on this dataset. While models based on neural networks are susceptible to overfitting when working with a small dataset (LIAR), they show great promise when working with a large dataset (Combined Corpus), as shown by a high F1-score.
Compared to traditional machine learning techniques, deep learning models are more effective at spotting fake news (Tables 3, 4). This contrast highlights the vulnerability of deep learning techniques to overfitting on smaller datasets, since it is more obvious on the large dataset Combined Corpus. Although it is a more traditional model, Naive Bayes (with n-gram) has tremendous promise in fake news recognition, coming very close to the performance of deep learning models and obtaining 93% accuracy on Combined Corpus. A study reveals that after around 2.5K training data, Naive Bayes' performance levels out and grows very slowly.

To put it another way, the rate of improvement of Bi-performance LSTMs as a deep learning model increases as the amount of the training dataset does (see Fig. 3). From this, it seems that, with enough training data, deep learning models can outperform Naive Bayes.

Revisiting the First Question: How well can classical and deep learning models discriminate between real and fake news? Overall, deep learning models perform better than their simpler counterparts. Deep learning will excel above traditional models if a big enough dataset is available. When given a short dataset, traditional models like Naive Bayes may perform very well, but deep learning algorithms have a tendency to overfit. To a greater extent as the dataset becomes larger, deep learning models outperform their more traditional counterparts. Could the most advanced pre-trained language models eventually outperform the tried-and-true models and deep learning models? (RQ2)

Table 5 displays the outcomes of several pre-trained language models on three separate datasets. These models may benefit from more complicated architectures, as opposed to deep learning models, which are prone to overfitting on a small dataset. This is because, with the exception of the final classification levels, pre-trained weights are used in all layers of these models. This allows them to refine their complex design without needing access to a large dataset. With an F1-score of at least 0.62 on the Liar dataset and at least 0.95 on the Fake or Real News dataset, all of the pre-trained models we examined outperformed the other standard ML and deep learning-based models. Due to the massive size of the dataset, pre-trained models like this perform better in the fake news detection task (i.e., Combined Corpus). It became out that among the pre-trained language models, the BERT and transformer-based models (i.e. BERT, RoBERTa, DistilBERT, ELECTRA) performed the best (i.e., ELMo). ELMo (93.6M parameters) only achieves 0.91 accuracy on the Combined Corpus dataset, while models like DistillBERT (66M parameters), BERT (110M parameters), Electra (110M parameters), and RoBERTa (125M parameters) all achieve 0.93, 0.95, and 0.96 accuracy, respectively. We have also seen that the more parameters a transformer-based model contains that have been pre-trained, the better it performs. This performance is backed by their state-of-the-art results on the text classification test (Liu et al., 2019; Sanh et al., 2019).

Revisiting the Second Question in Quick Order. Our findings demonstrate that, generally speaking, pre-trained models outperform both conventional and deep learning models, raising the question of whether or not state-of-the-art pre-trained language models are capable of surpassing these two types of models. Given that these models are pre-trained to learn contextual text representations on much larger quantities of text corpus and have produced new state-of-the-art in several text classification tasks, it is not surprising that they have outperformed traditional and deep learning models in the fake news detection task (Minaee et al., 2020).

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	.78	.76	.76	.93	.94	.93	.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**

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

the next step is to assess their performance. This comparison uses the Fake or Real News dataset because of the clear disparities in performance between the two teams. Their efficacy on randomly chosen training instances from the Fake or Real News dataset (n=500, 2500, and 5000) is shown. RoBERTa is shown to be far superior than the other two methods (Fig. 3). RoBERTa is able to reach over 90% accuracy with as little as 500 training data, and it improves with more samples. The accuracy increases to 98% with a sample size of 5000. Both Naive Bayes and Bi-LSTM have difficulties when the training dataset is tiny (less than 500 samples). However, even when using more data, they still can't achieve 90% accuracy with datasets smaller than 5000.
We also investigate RoBERTa's efficacy on sample sizes that are much smaller (Fig. 4). When the dataset size is increased to 300, we find that the model still has outstanding accuracy (84%). The reason for this is because RoBERTa's pre-trained weights have already mastered the semantic representation of massive text corpora. Corrections to the tagged headlines

Fig. 3. Analyzing the performance of Naive Bayes, Bi-LSTM, and RoBERTa on datasets of varying sizes (from Fake or Real News dataset).

4.3. Which model is the most effective when working with little data? (RQ3)

With little data, we discover that pre-trained BERT-based models may achieve impressive results. The fact that they outperform competing models even on very tiny datasets, such as Liar and Fake or Real News, is indicative of this. To double-check this, we choose the top performing model from the following three categories: Naive Bayes with n- gram (the gold standard), Bi-LSTM (deep learning), and RoBERTa (a hybrid of the two) (BERT-based)

Reading articles may help you develop the skills necessary to discern real news from fake news. When the sample size is decreased to fewer than 300, the model's performance quickly drops off. Because of insufficient information, the model cannot reliably differentiate between different types of news articles. Because of this, efficiency has plummeted.

Summary of RQ3. Which model performs best with small training data? Pre- In our experiment, trained models (i.e. RoBERTa) demonstrate respectable performance despite making do with a comparatively meager quantity of training data. With only 500 examples in the training set, RoBERTa is able to get an accuracy of over 90%. samples only (see Fig. 3).

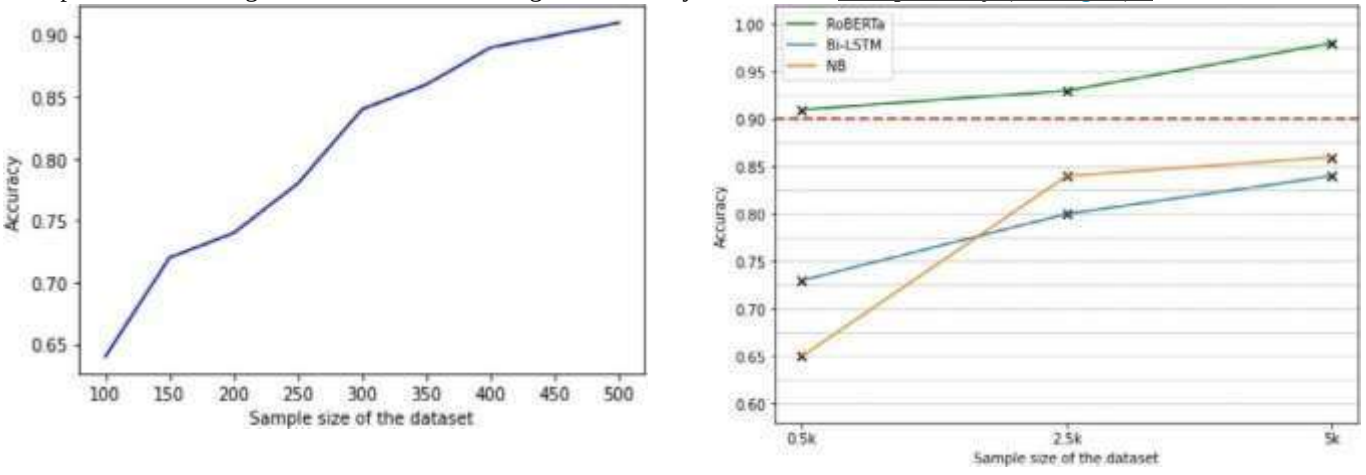

Fig. 4Examining RoBERTa's efficacy over a range of training dataset sizes (from Fake or Real News dataset).

Discussion
In this section, we compare the effectiveness of the 19 models we looked at across a range of parameters, such as feature set, resource requirements, etc (see Section 5.1). This section explains the findings from our investigation into the models' misclassifications. evaluating the effectiveness of different model types

Our final table of models, ranked by accuracy across all three datasets, is shown in Table 6. (Liar, Fake or Real, and Combined Corpus). The Naive Bayes model had the greatest accuracy across all three datasets we analyzed (pooled corpus (0.93), Fake or Real (0.86), and Liar 0.01). (0.80). (0.60). When we compared six common deep learning models across the three datasets, we discovered that three of them performed very well. Accuracy-wise, the best results are achieved by C-LSTM across the board (Accuracy = 0.95), HAN on Fake or Real (0.87), and HAN on the Liar dataset

(0.87). (0.90).

= 0.75). Across all three datasets we evaluated (combined corpus (Accuracy = 0.96), Fake or Real (0.98), and Liar (0.99)), RoBERTa was the top performer. This includes outperforming the other five pre-trained advanced natural language deep learning models we looked at. (0.62). On both the Combined corpus and the Fake or Real datasets, RoBERTa is the top performer across the models we considered, whereas HAN is the top performer just on the Liar dataset.

Performance-wise, Naive Bayes (with n-gram) is practically on par with that of deep learning and pre-trained language models. In light of this, Naive Bayes may prove to be an efficient technique for identifying fake news when applied to a suitably large dataset and limited by hardware. It has been found that Naive Bayes (with n-gram) is a good spam detector (Hovold, 2005). We find that on the Combined Corpus, the performance of Naive Bayes (with n-gram) is almost on par with that of deep learning models (see Ta- ble 3). If you don't have the hardware resources for deep learning and sophisticated pre-trained models (which may be the case for non-profit blogs/websites), Naive Bayes with n-gram might be a decent option with a large enough dataset. Bear in mind that the minimum required size of the dataset may be determined by the characteristics of the dataset, such as the number of topics. However, when trained on a small data set, Naive Bayes' accuracy suffers (see Fig. 3). When compared to other features when using the common learning models, bigram-based models (such Nave Bayes) perform well (lexical, sentiment, n-grams). Everywhere, the use of senti-There is a clear relationship between the length of the articles and the model findings (see Fig. 5).

vocabulary and vocabulary with emotion. Nothing in our research suggests that using Sentiment features might help you spot fake stories. Since lies may be spread on both sides of an issue, polarity (or emotion) has nothing to do with the veracity of a news item (positive or negative). Two LSTM variations (Bi-LSTM and C-LSTM) perform the poorest among the more traditional deep learning models when the dataset size is decreased (see RQ3). Improvements in LSTM-based models can be seen across the board, from the smaller LIAR dataset to the larger Combined Corpus dataset. Improving model performance and reducing the likelihood of overfitting may be accomplished by including more data in the original publication. As a result, neural network-based models may do well with datasets including more than 100,000 samples (Joulin et al., 2016).

Pre-trained BERT-based models outperform the alternatives across the board and in subsamples of the datasets (see RQ3). The BERT-based model (i.e., RoBERTa) achieves near-perfect accuracy (about 90%) with as little as 500 inputs. (Check out Fig. 3) These models may be useful for identifying fabricated news in several languages when a large quantity of labeled data is unavailable. Some examples of pre-trained BERT models for particular languages are ALBERTO for Italian (Polignano et al., 2019), AraBERT for Arabic (Antoun et al., 2020), and BanglaBERT.8.

We determined the average training time (per epoch) and GPU use for every BERT-based model on the Combined Corpus (during testing). When compared to BERT and RoBERTa, DistilBERT's training time is around half as long and it requires a smaller fraction of a GPU during testing (i.e. prediction) (see Table 7). DistilBERT may be appropriate for production-level use where hardware restrictions and decreased response times are a factor, since its 0.93 accuracy on the combined corpus is only slightly lower than BERT's (0.95 or RoBERTa's (0.96) 0.96). This is because DistilBERT is based on the principle of condensing information (Bucilua et al., 2006; Hinton et al., 2015). Its high efficiency and low material requirements make it suitable for use in manufacturing. Misclassification Analysis

Our evaluation utilizes three datasets, the best of which include pre-trained language models with accuracy levels of 96% or above on two of the three (Combined corpus and Fake or Real). An alternate data collection (Liar),

The models did not improve when we added progress indicators to them.

For example, there is no difference in performance when using a support vector machine (SVM) (0.71).

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				
SVMSVM LR				
Decision Tree AdaBoost Naïve Bayes				
These traditional models are used in different classification tasks includingtext classification. Different existing studies used them for fake news detection as well.				
Lexical 0.56	0.67	0.71		

	Lexical + Sentiment		0.56	0.66	0.71
	Lexical + Sentiment		0.56	0.67	0.76
	Lexical + Sentiment		0.51	0.65	0.67
	Lexical + Sentiment		0.56	0.72	0.74
Naïve Bayes	Unigram		0.60	0.82	0.91
	Bigram		0.60	0.86	0.93
k-NN	Empath		0.54	0.71	0.71

Deep learning models
CNN CNN extracts features and classifytexts by transforming words into vectors.
LSTM LSTM remembers information for long sentences.
Bi-LSTM Bi-LSTM analyzes a certain part from
both previous and next events.

C-LSTM Convolutional layer with max-pooling
combines the local features into a global vector to help LSTM remembering important information.
HAN HAN applies attention mechanism forboth word-level and sentence-level representation.
Conv-HAN Convolutional layer encodes
embedding into feature for word-level and sentence-level attention.
GloVe embedding

0.58 0.86 0.93

0.54 0.76 0.93

0.58 0.85 0.95

0.54 0.86 0.95

0.75 0.87 0.92

0.59 0.86 0.92

BERT These language models are~BERT embeddings 0.62 0.96 0.95
Advanced pre-trained language models
RoBERTa
DistilBERTELECTRA
pre-trained on large text corpus~ and can be fine-tuned for~ text classification.
RoBERTa embeddings 0.62 0.98 0.96
DistilBERT embeddings 0.60 0.95 0.93

ELECTRA embeddings 0.61 0.96 0.95ELMo ELMo embeddings 0.61 0.93 0.91

Table 7 ng time and GPU
Comparison usage (in testing) for BERT-based models. GPU used in testing
of traini #Parameters Avg. training time per epoch(GB)
Model (sec)

DistilBERT	66 M	2175	2.48
BERT	110 M	3149	2.95
RoBERTa	125 M	4020	3.07

The HAN model was the most accurate of the bunch (75%). When compared to the other two datasets, Liar's average article length (at 18 words) is much shorter (average 644 words for Combined Corpus and 765 words for Fake or Real news). We have shown that our model's performance improves when the average length of news stories increases, even while the amount of training data used stays the same (see Fig. 5). 5,000 records were selected at random from each of the three datasets, and they were all evaluated using the Naive Bayes model to ensure accuracy. Similar results have been found with other models. Because more information can be collected from lengthier news stories, they tend to provide more accurate models.
While Bogus or Real News and Liar both provide examples of false news on political issues, the third database, Combined Corpus, has examples of fake news on a broad variety of topics, such as health, research, politics, the economy, and more (see Fig. 1 in Section 3). With the intention of discovering whether or not the topic of the news affects the classification, we employ topic-based analysis on the fake news items from the Combined corpus that our model incorrectly labels as genuine. We then use the 10 overarching themes from the combined corpus as a road map to determine how each misclassified example fits into the bigger picture (Fig. 1). Misquoted statements are frequently used and misconstrued in fake news. More commonly than any other words,

the articles include the phrases "said," "study," and "research." A red indicator indicating the source might be a fraud is an abundance of the word "said."

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

to use quotes to support their own arguments and make them look more compelling.

Using a topic-by-topic analysis of misclassification in the combined corpus, we discover that 49.6% of false positive news (articles mistakenly labeled as fake) pertain to health and scientific research (Table 8). Alternatively, politics is seldom included in false positive news reports (just 27.6% of them). In view of the aforementioned high false positive rate, there is much room for improvement in the creation of convincing health and research-related clickbait news. As long as the original study piece isn't drastically changed, the false news will be almost indistinguishable from the real thing. This facilitates the dissemination of false information about the development of a vaccine for deadly diseases like cancer on clickbait news sites. Since then, the media has made significant investments to counter the spread of disinformation.

As with political news, it is crucial that media outlets work to limit the dissemination of false information in the fields of medicine and science to protect the public's health and safety. If you need an example, just think about how much of an impact fake news has had on the current COVID-19 epidemic. Inaccurate information about the corona has caused confusion and distress among a large audience. The effects of false claims like "Alcohol heals COVID-19" and "5G spreads coronavirus" may be seen in the actual world. 9 Due to the severity of the situation, fake corona reports have been likened to a new pandemic or infodemic. 10

Conclusions

Thus, we provide a thorough analysis of the efficiency of 19 different machine learning approaches, comparing their results over three separate datasets. Eight of the models are deep learning engines, six are conventional learning engines, and five are cutting-edge pre-trained language models like BERT. According to our findings, BERT-based models are superior to their counterparts on all datasets. Moreover, our results demonstrate that pre-trained BERT-based models may sometimes beat their untrained counterparts on small sample sizes, implying that their performance is robust regardless of the size of the dataset. We also find that Naive Bayes with n-gram may get results that are competitive with neural network-based models when the dataset size is sufficiently big. Both the size of the dataset and the level of information in a news item greatly affect the performance of LSTM-based models. When there is enough information in a news item, LSTM-based models are more likely to be able to avoid overfitting. Our comparative research and its findings will help educate future studies in this field and enable companies (such as online news portals and social media) in choosing the optimal model for recognizing fake news. We will be devoting a lot of time and energy to creating algorithms that can detect the predominance of misinformation and health-related false news that spreads on social media during the current COVID-19 outbreak.

Certificate of Responsibility in Editing and Publishing (CRediT)

Junaed Younus Khan revised the methodology and the circuitry. M. Tawkat Islam Khondaker, the methodology, and the writing (review and editing) processes are all discussed. The multitalented Sadia Afroz has worked as a writer, editor, and proofreader. Methodology, Writing, Review, and Editing by Gias Uddin. Anindya Iqbal's Methodology, Writing Review, and Editing.

Recital of Conflicting Interests

All authors have confirmed that there are no financial or personal ties between them that may be interpreted as having affected the work disclosed in this publication.

References

Adhikari, Ram, Tang, and J. Lin conducted the research (2019). Bert, as Docbert, is the document management software you need. You may download the PDF version of this paper at arXiv:1904.08398.

Drs. H. Ahmed, I. Traore, and S. Saad (2017). Use of N-grams and machine learning to identify fake news stories online. Proceedings of the International Conference on Intelligent, Secure, and Reliable Systems in Distributed and Cloud Environments (pp. 127–138).

Springer.

Authors: H. Allcott and M. Gentzkow (2017). A critical analysis of how social media and fake news affected the outcome of the 2016 presidential election.

Journal of Economics and Societies, 31(2), 211-236. Economic Perspectives.

Reference: Antoun, W., F. Baly, and H. Hajj (2020). Using a Transformer-based Model for Arabic Language Processing (Arabic: AraBERT). To see the preprint, visit arXiv:2003.00104.

Follow this citation format: Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. To be found in Research in Machine Learning: Volume 3, Issue 1 (January), Pages 993-1022.

Specifically, Bondielli, A., and F. Marcelloni's work is cited (2019). Techniques for spotting fake news and other forms of

propaganda. The reference is to Information Sciences, 497, 38-55.

9 Visit https://www.bbc.com/news/stories-52731624 for more details (last visited October 5, 2020). This item was retrieved on October 5, 2020, from https://www.nature.com/articles/d41586-020-01409-2.

"Rehm, G., P. Bourgonje, and J. M. Schneider" (2017). A technique for recognizing the inherent bias in titles and abstracts, which aids in the detection of clickbait and fabricated news. To appear in the proceedings of the 2017 EMNLP Workshop on Natural Language Processing and Journalism (pp. 84–89).
Citation Information: Bucilua, C.; Caruana, R.; Niculescu-Mizil, A. (2006). The model's size has been reduced. Knowledge Discovery and Data Mining: The Proceedings of the 12th Annual ACM SIGKDD Conference (pp. 535–541).
Carroll, John D., and Paul Arabie (1998). Multiple-dimensional scaling. Evaluating, judging, and selecting (pp. 179–250). Elsevier.
Authors Y. Chen, N. J. Conroy, and V. L. Rubin contributed to its creation (2015). Internet misinformation: recognizing propaganda in the form of clickbait. Workshop on Multimodal Detection of Deceit 2015: Proceedings from the Association for Computing Machinery (pp. 15–19). ACM.
The authors (Le, Q. V., Clark, K., Luong, M.-T., and Manning, C. D.) (2020). In order to make text encoders more effective as discriminators than generators, "Electra" pre-trains them. This study was previously available on arXiv with the working title 2003.10555.
Modest Proposal: A Cliché by M. (2014). An Detector of Sarcasm Read more about the Sarcasm Detector here: http://www.thesarcasmdetector.com/.
The authors N. J. Conroy, V. L. Rubin, and Y. Chen (2015). Dishonesty detection algorithms; fake news debunking approaches. Proceedings of the 78th Annual Conference of the Association for Information Science and Technology: Information science with impact: Research in and for the community (p. 82). The American Society for Information Science and Technology.
Dai, E.; Sun, Y.; Wang, S. (2020). Contrary to popular belief, ginger does not have anti-cancer properties. Proceedings of the 14th International AAAI Conference on Web and Social Media (pp. 853–862).
the authors van Noord, G., de Vries, W., A. van Cranenburgh, A. Bisazza, T. Caselli, and M. Nissim (2019). Bertje exemplifies the kind of Dutch bert you'd expect to find. The preprint may be found at arXiv:1912.09582.
The authors of this paper are Lee, K., Toutanova, K., Devlin, J., and Chang, M.-W. (2018). The "foresightful pre-training of deep bidirectional transformers" for language understanding, to use a phrase from Bert. Document ID: 1810.04805 on the arXiv preprint server.
Authors: Dwivedi, S. M., and S. B. Wankhade (2020). Research on techniques for spotting fake news. Image Processing and Capsule Networks International Conference Presentation (pp. 342–348). Specifically, Fast E., Chen B., and Bernstein MS (2016). Recognizing Important Concepts Among a Lot of Text is a Sign of Emotional Intelligence (EQ). Proceedings of the 2016 ACM Conference on Human Factors in Computing Systems (pp. 4647–4657). Choi Y., Feng S., and Banerjee R. (2019). (2012). Syntactic stylometry for spotting lies. Abstracts of selected lightning lectures given during the 50th annual meeting of the Association for Computational Linguistics (pp. 171–175). The initials stand for the Association for Computational Linguistics.
Research by J. (1998). In this study, we used N-gram features to the task of text categorization. Journal of Austrian Computer Science, Vol. 3, No. 1 (1998): 1–10.
As an additional example, Gilda, S. (2017). There is now an investigation on the efficacy of several machine learning-based strategies for identifying instances of fake news. 2017 IEEE's 15th Annual Student Conference on Research and Development (with Points) (pp. 110–115). Sigmundo González-Carvajal and Edgardo C. Garrido-Merchán, Institute of Electrical and Electronics Engineers (2020). Comparing BERT's text classification results against those of more traditional ML techniques. There is a preview available at arXiv:2005.13012.
V. Mesyura & M. Granik (2017). Exposing hoaxes using a naive Bayes classifier. This year's IEEE Ukraine Conference on Electrical and Computer Engineering (UKRCON 2017) (pp. 900–903). Abstract of the IEEE article by Gravanis, G.; Vakali, A.; Diamantaras, K.; and Karadais, P. (2019). Here are the signs: Methods for detecting fake news are compared and contrasted. Expert 129, Systems with Applications, pages 201-213.
Y. Slimani, T. Hamdi, H. Slimi, and I. Bounhas (2020). This technique is able to go through all the tweets and spot the fake news by combining user attributes with graph embeddings. Books: Distributed Computing and Internet Technology: Proceedings of the International Conference (pp. 266–280). Springer. According to Hinton (author), Vinyals (author), and Dean (author) (2015). Using a neural network to condense the data.
see the preprint at arXiv:1503.02531.
Hovold, J. (2005). Naive Bayes algorithm-based spam detection using word-position features. Semester shared between Europe and Asia
Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. Equipment for properly tagging texts. ArXiv:1607.01759 has a PDF version that you may download.
Oh, The authors (D. Jwa, K. Park, J. M. Kang, and H. Lim) (2019). Science Applications, Vol. 9, No. 19 (2017): exBAKE: A Bidirectional Encoder-Based Model for Automatic Fake News Detection (4062).
From what I can tell, Khattar (D), Goud (J. S.), Gupta (M), and Varma (V). Here, we introduce a multimodal variational autoencoder (Mvae) for spotting fake news stories. As part of the conference call
Kim, Y. (2014). (2014). (2014). Sentence classification using convolutional neural networks. For more reading, please visit

arXiv:1408.5882.
Authors: S. Kula, M. Chora, and R. Kozik (2020). The BERT-based architecture was used to detect hoaxes. Springer's Conference on Complex, Intelligent, and Software-Intensive Systems.
We thank U. Lall and A. Sharma for their contributions (1996). Re-sampling hydrologic time series using a nearest-neighbor bootstrap. This article may be found in volume 32, issue 3 of Water Research, pages 679-693.
Lee, Z. Liu, and P. Fung are the authors (2019). Team yeon-de-noising zi's of poorly-labeled data to discover hyperpartisan news won Task 4 at Semeval-2019. The Thirteenth International Workshop on Semantic Evaluation Proceedings
The authors are credited as follows: D. Leonhardt and S. A. Thompson. It's no secret that Trump lies. The newest issue of The New York Times is the twentieth.
Works by Li, X.; Bing, L.; Zhang, W.; and Lam, W. Sentiment analysis that is both thorough and based on several points of view may be performed with the help of BERT. See the preprint at arXiv:1910.00883.
Lin, J. (1991). (1991). (1991). Divergence was calculated using the Shannon entropy. IEEE Trans. Inf. Theor., 37(1), pp. 145–151.

Liu, Y. (2019). (2019). Adjust BERT so that it can provide an accurate summary. Refer to the preprint at arXiv: 1903.10318.
Researchers Levy, Lewis, Zettle- moyer, Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., and Chen, D. (2019). Method for Robustly Optimizing Bert Pretraining (or Roberta for short). You may find a pre-print of this work at arXiv:1907.11692.
The authors of this study are cited as Loshchilov (I.) and Hutter (F. (2017). Weight loss regularization that is not related to dieting. Its PDF version may be found at arXiv:1711.05101.

Those responsible for this work include S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao (2020). The use of deep learning for text classification is analyzed in detail. To see the preprint, visit arXiv:2004.03705.
The authors are Munikar, S. Shakya, and A. Shrestha (2019). Using bert to accurately categorize people's feelings. In 2019, we'll see the publishing of the first edition of Artificial Intelligence: Its Business and Social Impact (AITB). (pp. 1–5). In a paper published in IEEE, Oshikawa, Qian, and Wang (W. Y. (2018). The current status of NLP-based fake news detection is analyzed. There's a downloadable PDF at arXiv:1811.00770.
Peng, Yan, and Lu, all named Y. For the purpose of determining whether method is best for transfer learning in biomedical natural language processing, we evaluate bert and elmo on 10 gold-standard datasets. You may get the preprint at arXiv:1906.05474.
John Pennington; Robert Socher; Christopher Manning. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): "Word Representation Using Global Vectors (Glove)"
Authors: Peters, M. E.; Neumann; Iyyer; Gardner; Clark; Lee; and Zettlemoyer, L. It is the context in which a word is used that gives it its meaning. Please see arXiv:1802. 05365 for the preprint.
Researchers showed that the risk of developing type 2 diabetes was lowered when people ate a diet high in both dietary fiber and protein (Polignano, Basile, de Gemmis, Semeraro, and Basile, all 2013). AlBERTo, a rival system for NLP that processes tweets, using the BERT model of language comprehension. Aiming at CLiC-it.
Our gratitude goes to L. Prechelt for his contribution here. Applying cross-validation to determine when to stop utilizing quantitative criteria. 11(4), 761-767, Artificial Neural Systems.
The work of L. Prechelt. When it's OK to end things before they're over. Strategies for Neural Networks in the Real World (pp. 55–69). Springer.
The authors of the study are Y. Choi, H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. There is complexity in the truth: understanding the vocabulary of alternative media and political fact-checking. L. Rosasco; E. D. Vito; A. Caponnetto; M. Piana; and A. Verri. (2017). Proceedings of the 2017 Empirical Methods in Natural Language Processing Conference. It is reasonable to suppose that all loss functions are the same, right? The 16th edition of Neural Computation is between 1063 and 1076 pages lo ng.
Three authors: Victor L. Rubin, Yichen Chen, and Nicholas J. Conroy. The three most common methods for identifying fake news. To Hold Its 78th Annual Gathering Information Science and Technology Association Proceedings: An Application-Oriented Approach to Information Science and Community-Based Research The American Society for Information Science and Technology.

Published by Chen, Y., Cornwell, S., Conroy, N., and Rubin, V. Is this an elaborate fraud, or the genuine deal? Identifying Fake News Using Satirical Indicators. Proceedings of a Computational Workshop on Detecting Deception
This is the work of N. Ruchansky, S. Seo, and Y. Liu. In order to spot fake news, the Csi team created a hybrid deep learning model. Knowledge and Information Management: Proceedings of the 2017 ACM International Conference on.
It was written by T. Wolf, J. Chaumond, L. Debut, and V. Sanh. Distil is more compact, swifter, less expensive, and lighter than BERT. The well-known BERT has been condensed into BERT. See the preprint at arXiv:1910.01108 for further information.
abilities to defend: identifying and explaining bogus news. Liu, H.; Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. Presented at the 25th Annual Meeting of the ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining), San Francisco, CA, August 2010
By Shu, Sliva, Wang, Tang, and Liu: Data mining to detect social media hoaxes. This refers to the ACM SIGKDD Explorations Newsletter, volume 19, number 3, pages 22–36.
Singhania, Fernandez, and Rao each contributed to the book. The Adaptive Network on the Third Horizon (or 3HAN): Fake-news detection using a deep neural network. International Conference on Neural Information Processing, Springer, pp. 572–581.
Authors: Xu, Y., Sun, C., Qiu, X., and Huang, X. If I want to use bert for text labeling, what should I do to make it work as

efficiently as possible? Recently, China hosted the Annual National Conference on Chinese Computational Linguistics. Springer.
Co-authors: Moret, S.; de Alfaro, L.; Ballarin, G.; Tacchini, E.; Della Vedova, M. L. It's true that some individuals take great pleasure in immediately spotting hoaxes in their social media feeds. Please see the preprint at arXiv:1704.07506 for details.
Ingrid Tenney, Daniel Das, and Elizabeth Pavlick. BERT rediscovers the classic NLP pipeline phases. The preprint may be accessed at arXiv:1905.05950.
Authors: John Thorne, Ming Chen, George Myrianthous, Jian Pu, Xiaobo Wang, Antonis Vlachos. Detecting the slant of bogus news stories by using a stacked ensemble of classifiers. To appear in the proceedings of the 2017 EMNLP Workshop on Natural Language Processing and Journalism (pp. 80–83).
The name is W. Y. Wang. "lie, liar, pantless liar": presenting a new standard dataset for spotting fake news. Obtainable as a PDF file at arXiv:1705.00648.
Xu Hu, Hong Liu, Luo Wu, and Jiao Liu. Apply your newfound knowledge of past online rumors to the detection of today's spreading urban legends. Information Analysis and Mining: Papers from the 2017 SIAM International Conference on Information Analysis and Machine Learning.
The authors are L. Wu and H. Liu. Tracing the digital footprints of bogus news: analyzing the features shared by popular online content. Proceedings of the Eleventh Annual ACM International Conference on Web Search and Data Mining. ACM.
It should be noted that the paper was written by Zhang X. and Ghorbani A. A. This summary discusses what fake news is, how to spot it, and how to talk about it. 57(2) of the journal Information Processing & Management contains article 102025.
Xiaozhou Zhou and Roberto Zafarani. In order to spot hoaxes that have been widely shared online, a pattern recognition algorithm is being used. Publication information: Explorations, 21(2), 48-60 (Newsletter of the ACM Special Interest Group on Knowledge Discovery in Databases).