Fake News Detection by Fine Tuning of Bidirectional Encoder Representations from Transformers

MSVPJ Sathvik, *Member, IEEE*, Kanishk Bajpai, Prashant Kumar Singh, Boddu Moses Vijay Kumar, Mukesh Kumar Mishra, Sibasankar Padhy,

***Abstract*—Everyone now has internet and social media access, making it simple to get information and news. On the other hand, there are fake news articles, also. It not only makes difficult for the public to find their truthfulness but also misleads them. Consequently, developing intelligent systems for separating news is critical. In this paper, four deep learning techniques such as Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Networks-BiLSTM (CNN-BiLSTM) are used to detect fake news. We trained these models on three large datasets, namely, CoAID, GossipCop, and PolitiFact, that include fake news from social media and news articles. The text is divided into two categories: fake and real. We also trained and compared the results with ten well-known machine-learning classifiers, e.g., Naive Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbors Classifier, etc. The experimental results indicate that the deep learning-based BERT method achieves better accuracy and F1-score of 91.94% and 92.06% for PolitiFact, 99.09% and 98.38% for CoAID, 85.59% and 82.40% for GossipCop, respectively. These results suggest that deep learning algorithms provide favorably accurate results and outperform machine learning models in terms of accuracy.**

***Index Terms*—Social Media, Fake News, Machine Learning, Deep Learning, BERT.**

I. Introduction

In this modern but still developing world, several essential things have been achieved with the new advancement of technology updating every day. One of them is social media being compelling in its current state. Nowadays, news plays an essential role in our lives, keeping us informed and updated about the incidents happening nearby and beyond our immediate environment. Due to social media and the internet, the news is at everyone's fingertips. However, these days every piece of shared information should not be trusted blindly, at least not without checking correctly. People manipulate mass media to make money by distorting information in many ways. They use this power to spread fabricated articles to bias and alter their opinion, do scams and various illegal activities to gain readership, financial or political gain, and violate a group, person, or agency. Fake News is news reports that are intentionally and verifiably false and misleading. Such news is deliberately created articles from satirical websites. The misleading, false news outwits the more trustworthy, costly, time-consuming, comparably lower reachability and spreading power of conventional news due to its low cost, reduced time consumption, and improved reachability. Fake news needs to be persuasive, e.g.,

i **Financial (Share Markets):** In 2013, false information about Barack Obama's injuries in an explosion was disseminated on Twitter, resulting in the loss of 130 billion dollars in market value [1].

ii **Election :** According to a survey [2], 64% of US people said false news had created a significant lot of doubt regarding the integrity of reported events. Likewise, during the 2016 US presidential campaign, false news was accused of being the primary contributor to increased political division and party battles and influencing the outcome.

Literature believes the 18th century to be the official birth date of fake news because it was misleading and compelling [3]. In 1779, Benjamin Franklin sent a complete forged letter addressed to Capt. Samuel Gerrish. It was designed to appear as a regular supplement to a Boston newspaper, detailing the atrocities the British and their allies committed. He was explicitly attempting to sway public opinion when the peace talks began.

When Tim Berners Lee created the first web page in 1991, the world wide web became publicly available, and the genesis of social media websites appeared in 1997. In 2012, only 49% of individuals reported viewing news on social media [4]. However, by the end of 2016, about 70% of the population was reliant on social media. In 2017, Tim Berners Lee stated that false news was one of the three most critical and unsettling Internet phenomena that needed to be addressed before the Internet could fully serve mankind [5].

Fake news may be classified broadly into three categories [6]:

i) **Serious Fabrications:** Most falsified articles or news are created for the reporter's self-promotion above public information validity, boosting readership or viewership. They create deceptive headlines to get people to click and fraudulent reasons, e.g., extortion, defamation, and installation hale.

MSVPJ Sathvik, Boddu Moses Vijay Kumar, Prashant Kumar Singh, Kanishk Bajpai, Mukesh Kumar Mishra, and Sibasankar Padhy are with the Department of Electronics and Communication Engineering, Indian Institute of Information Technology, Dharwad, Karnataka, 580009, India. e-mail:20bec024@iiitdwd.ac.in, 20bec010@iiitdwd.ac.in, 20bec033@iiitdwd.ac.in,20bec020@iiitdwd.ac.in,mukeshkumar@iiitdwd.ac.in, sibasankar@iiitdwd.ac.in.

Manuscript received Feb xx, 2023; revised August xx, xxxx.

ii) **Large-scale Hoaxes:** Deliberate fabrications that go beyond pranks or jokes in an attempt to deceive spectators. Viewers who become sufferers may suffer financial loss or be physically or emotionally harmed.

iii) **Humorous Fakes:** Fabrications often provided in the manner of professional journalism appear to be trustworthy news, but with a clue to alert readers to the story's hilarious nature or goal.

A. Related Work

As discussed above, much fake news is being uploaded on social media; thus, it is very challenging to inspect whether every piece of information is accurate. Hence, the development of automated fake news detection techniques will be beneficial. Recent research suggests that machine learning and deep learning algorithms may play a vital role in identifying fake news. Besides, machine learning algorithms are readily scalable and can be tuned to low complexity. Barbara Probierza et al. explored various machine learning algorithms, namely random forests, support vector machine (SVM), AdaBoost, etc., to detect fake news [7]. They reported that the SVM approach achieved the best results with 94.19% accuracy. Similarly, SVM classifier has been examined for predicting fake news using dimensionality reduction in [8].

In [9], authors investigated various learning techniques, e.g., logistic regression (LR), Naive Bayes (NB), and random forest. They found that logistic regression had the best results of 98.93% as compared to others. Agarwal et al., [10] explored learning techniques such as SVM, NB, KNN, convolutional neural networks (CNN), and LSTM for detecting fake news. Furthermore, they recorded 97% accuracy for LSTM method. In [11], authors compared machine learning algorithms with deep learning algorithms. They included NB, KNN, artificial neural network, and LSTM for the experiment and reported that LSTM outperformed all the techniques and achieved 94.21% accuracy. In [12], authors executed bidirectional encoder representations from transformers (BERT), LSTM, CNN, NB, passive aggressive classifier (PAC), gradient boosting (GB), decision tree (DT), etc. Further, numerically they demonstrated that BERT performed better with 58.8% accuracy. Raza et al. [13] executed various popular transformer pre-trained models for predicting fake news. They studied BERT-u, BERT-c, GPT-2, and different popular transformer models and gained an accuracy of 74.8%. In [14], authors examined generative adversarial neural networks (GAN), BERT, and other popular deep-learning methods. In [15], authors implemented distilBERT and compared the metrics with different deep learning algorithms. They noted an accuracy of 97.20%. In [16], authors executed FDNet neural network with Glove embeddings and significantly improved the performance. The authors reported an accuracy of 98.36%. In [17], authors implemented six baseline machine learning models and two deep learning models. They examined LSTM and GRU with PolitiFact, GossipCop, and covid-19 fake news dataset. In [18], CNN performed with TF-IDF vectorization, and the F1-score of 96.89% was noted. In [19], Random Forest was executed for fake news prediction and recorded an accuracy of 84.40%. In [20], Recurrent neural networks and transformer models are executed for fake news prediction and examined with other machine learning and deep learning models. An accuracy of 84.40% was reported in work. Authors of [21] executed Bi-LSTM for fake news detection and studied different metrics in the experiment. They found an accuracy of 90.40%.

Fake news detection is a natural language processing binary classification. Until now, no algorithm is perfect, i.e., no algorithm yet reported 100% accuracy. Accordingly, researchers are trying to improve accuracy by developing better models to identify fake news and information. Keeping these challenges in mind and the capability of learning algorithms to identify fake news and articles, this paper aims to investigate the different deep learning algorithms to detect fake news on three datasets GossipCop, CoAID, and PolitiFact. We also aim to raise awareness about fake news and articles faced by social media and the internet. The contributions of the work are summarised as follows:

- This work fine-tunes the BERT for fake news prediction and develops a framework to detect fake news that performs satisfactorily on several real-world datasets, namely PolitiFact, CoAID, and GossipCop. The performance of the models is evaluated in terms of accuracy, precision, recall, F1-score, and confusion matrix.
- We also perform experiments to examine the accuracy of BERT with leading deep learning methods, for instance, LSTM, BiLSTM, CNN-BiLSTM, and ten well-known machine learning techniques such as PAC, LR, Multinomial NB, DT, KNN Classifier, SVM, GB, etc., and some useful insights are drawn related to the fake news detection.
- Furthermore, the accuracy of the BERT is also compared with the existing work reported in the literature on fake data detection. Experiment results indicate that the proposed method achieves an accuracy of 99.08% on the CoAID dataset (News related to Covid-19), 91.94% on the PolitiFact dataset (US political news), 85.59% on the GossipCop dataset.

The paper is structured as follows: Section II introduces the proposed methodology. The experiment results and discussion are examined in section III. Finally, conclusions are presented in section V.

II. Methodology

The models' training is divided into two categories. The first category is typical machine learning model training. For instance, the machine learning models are passive-aggressive, decision tree, MLP classifier, multinomial naive Bayes, logistic regression, random forest classifier, K neighbors classifier, SVM, adaptive boosting, and gradient boosting. The second group includes deep learning training techniques, e.g., LSTM, Bi-LSTM, CNN-BiLSTM, and BERT. There are multiple phases to the training. The first phase is pre-processing, which involves removing stop words, extraneous data, lemmatization, tokenization, and lowercasing. The second step is to divide the data into training and testing sets. The third phase is feature extraction. We use Term frequency–inverse document frequency (TF-IDF) to extract features for classical machine learning models, whereas word embedding is used for deep learning models. In the fourth stage, we create the neural network architecture. The models are analyzed using the metrics such as accuracy, precision, recall, F1-score, support, and confusion matrix.

A. Data Collection

The training of the models is conducted with three fake news datasets, namely, GossipCop, PolitiFact, and CoAID (COVID-19 healthcare misinformation dataset) [22]. GossipCop and PolitiFact are collected from FakeNewsNet [23].

The GossipCop Dataset: The GossipCop contains news regarding gossip on celebrities. This dataset has two files gossipcopreal.csv and gossipcopfake.csv. The first file gossipcopreal.csv contains tweets of correct information, over 5328 tweets, and gossipcopfake.csv file includes tweets of fake news, over 5322 tweets. Both files are merged and labelled real news as 1 and fake news as 0. The dataset contains columns naming id, URL, title, tweet-id, and true. We have used the title as text [23].

The PolitiFact Dataset: PolitiFact dataset contains the news and tweets regarding US political news. The dataset has two files i) politifactreal.csv, which includes 432 real news tweets, and ii) politifactfake.csv has 618 fake news tweets. In this case also, both files are merged and labelled real news as 1 and fake news as 0. The dataset contains columns naming id, url, title, tweet id, and true [23].

CoAID (COVID-19 healthcare Dataset): It is a dataset of tweets and articles circulated over Covid-19. The dataset contains 5486 tweets and news articles regarding Covid-19. The dataset includes posts related to Covid-19, tweets, and replies as well. The topics in the dataset include COVID-19, coronavirus, pneumonia, flu, lockdown, stay home, quarantine, and ventilator. The claiming is done by referring to WHO official website, Twitter account, and MNT [22].

B. Data Preprocessing

Data preprocessing is crucial for training in neural networks and machine learning models. The tweets have data that is unstructured. Deep learning and machine learning algorithms need structured data for the best performance. The raw data has to be structured and refined to make it machine-readable.

- **Lower Casing:** It converts capital letters into small letters, e.g., ‘NLP’ to ‘nlp’
- **Removing Unimportant Data:** Removing insignificant data which are not required, like full stops, question marks, commas, etc.
- **Tokenization:** It is the process of segregating words from sentences, for instance, ‘I am a student’ to ‘I’ ‘am’ ‘a’ ‘student’.
- **Removal of Stop Word:** Stop words are the most repetitive words and have no significance, like conjunctions, prepositions, etc. These types of words are removed from all three datasets.
- **Lemmatization:** It is used to reduce the actual word to root word to avoid the excess of word types. For example: ‘ate’, ‘eaten’, ’eating’ to eat.

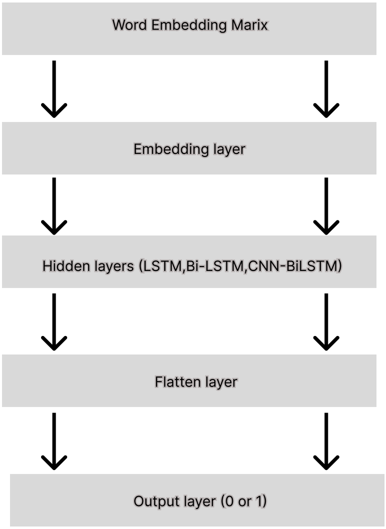


Fig. 1: The figure shows the deep neural network architecture

C. Data Splitting

During data splitting, 90% of the data are utilized for training and 10% for testing. The training data is used to train for the machine learning model’s input and output. The training data is the unseen data for the trained model and is applied to evaluate the model.

D. Learning Techniques

1) Deep Learning Model: This work examines four deep learning models. This is done in two steps. The first step is to extract features from the text using a deep learning feature extraction method.

Feature Extraction: The neural networks employed in deep learning models are trained only on numbers and float values. The neuron cannot take direct text as input. Therefore, it is essential to convert them into the form of vectors. We convert the natural language to machine-understandable language. In this work, the Gensim library is used for this purpose.

Deep Neural Network: Fig. 1 demonstrates the functional architecture of experiments applied to detect fake news. In this figure, the embedded matrix represents the features of the news in the dataset, and word embedding matrix is embedded into the embedding layer and then into the hidden layer. The hidden layer is the core of the entire experiment. The most critical operations are done in the hidden layers.

Long Short Term Memory (LSTM): LSTM has a memory block and three multiplicative gating units. It is capable of memorizing the previous sentences and previous text. The neural network, built on dense neurons, cannot memorize the previous sentences. Fig. 2 and Fig. 3 depict the working principle and different gates in the LSTM layer, respectively.

Forget Gate Layer: g_t, as shown in the figure, decides which information to exclude from the cell state [17].

$$g_t = \sigma(W_g[H_{t-1}, X_t] + b_g)$$

The new information is added at input gate layer n_t as given in the equation

$$n_t = \sigma(W_n[H_{t-1}, X_t] + b_n)$$

Then decides which values to get updated

$$S'_t = \tanh(W_s[H_{t-1}, X_t] + b_s)$$

Now, the cell has to be updated. Updating the cell follows the below mathematical expression

$$S_t = g_t \times S_{t-1} + i_t \times S'_t$$

Then the weights are forwarded to output gate layer. The output o_t follows the equation

$$o_t = \sigma(W_o[H_{t-1}, X_t] + b_o)$$

After that, it is forwarded to $tanh$ layer so that the values range from -1 to +1.

$$H_t = o_t \times \tanh(S_t)$$

Peephole: Let the gate layer look at the cell state as illustrated in the following equations

$$g_t = \sigma W_g[S_{t-1}, H_{t-1}, X_t] + b_g)$$

$$n_t = \sigma(W_n[S_{t-1}, H_{t-1}, X_t] + b_n)$$

$$o_t = \sigma(W_o[S_t, H_{t-1}, X_t] + b_o)$$

Coupling forgot, and input gates follow the equation

$$S_t = g_t \times S_{t-1} + (1 - g_t) \times S'$$

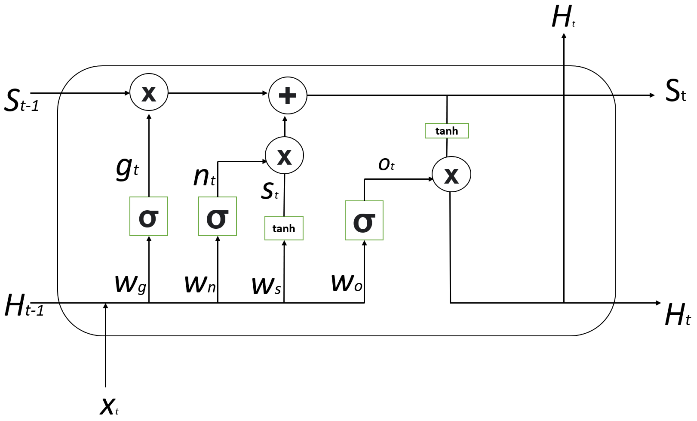


Fig. 2: LSTM (Long Short Term Memory) [17]

BiLSTM: BiLSTM is a deep-learning neural network that can operate past and future features of the text. It has the ability to understand the previous words and future words and may differentiate accordingly. The BiLSTM contains two LSTM layers and is interconnected to each other to extract features of the upcoming words and previous words. One LSTM layer is for future features, and the other is for extracting past features (as demonstrated in Fig. 4) [24].

CNN-BiLSTM: The convolutional neural network (CNN) is popular for extracting features from text and images. CNNs have significant importance in computer vision. They are used for most of the computer vision applications, from simple applications to generative adversarial neural(GAN) networks. Researches also show that CNNs perform well on text as well. In this neural network, the text is passed into CNNs to extract the features and then passed into the BiLSTM layer. The input of BiLSTM has extracted features from CNN layers, as shown in Fig. 5 [25]. The below mathematical equations describe the working principle of CNN [26].

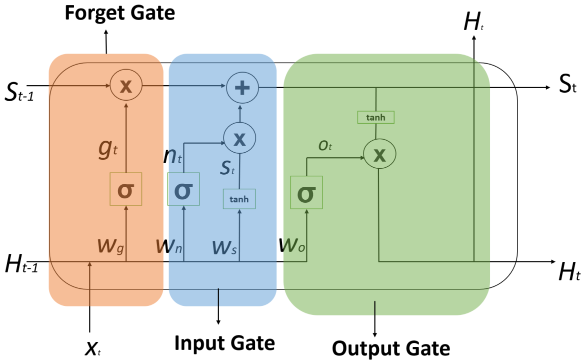


Fig. 3: Figure demonstrates the different layers in LSTM

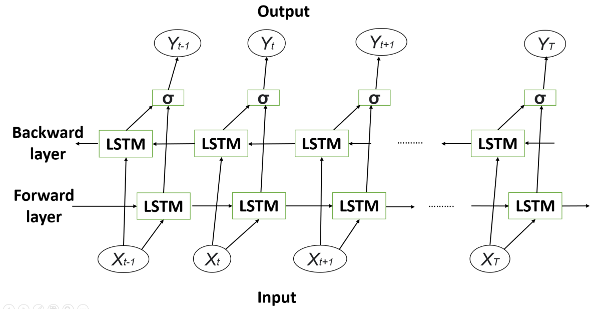


Fig. 4: Bi-LSTM [24]

$$\text{Convolution :} Z^l = H^{l-1} * W^l$$

$$\text{Max Pooling :} H^l_{XY} = \max_{\text{i=0...s, j=0...s}} H^{l-1}\text{(X+i)(Y+j)}$$

$$\text{Fully-connected Layer :} Z_l = W_l * H_{l-1}$$

$$\text{ReLu (Rectifier) :} ReLU(Z_i) = max(0, Z_i)$$

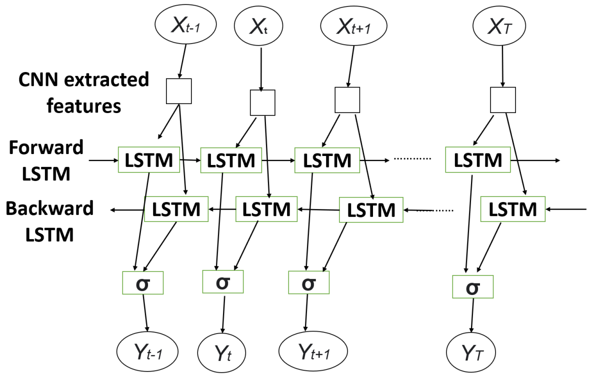


Fig. 5: CNN BiLSTM [25]

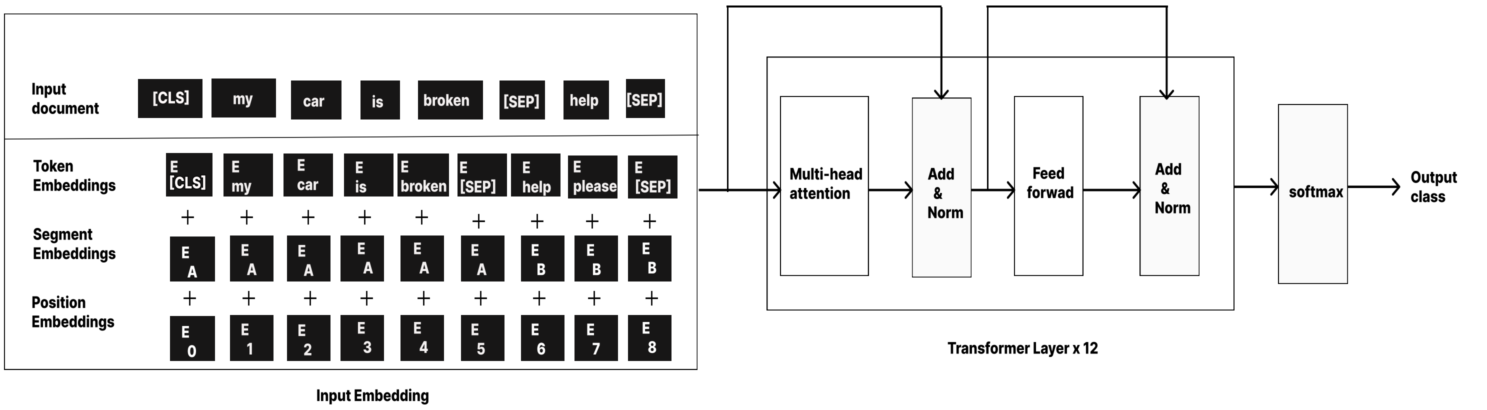


Fig. 6: An illustration of the working of BERT

BERT: BERT stands for bidirectional encoder representations from transformers. It was introduced and developed by Google. It outperformed most NLP applications like machine translation, sentiment classification, next-sentence prediction, etc. It can absorb the features of the previous and upcoming words at once. As it is bi-directional, it understands language better than other models. In this work, we use BERT base uncased, pre-trained on the English language using masked language modeling (MLM). The training of the BERT involves fine-tuning of BERT base uncased. Fig.6. illustrates the working of BERT [27]. The attention is mapping a query and key-value pairs as an output, which includes all of those vectors. The result is a weighted sum, with the weight allocated to each value calculated by a compatibility function of the query with the relevant key. Quotations are defined in [28]

$$\text{Attention}(Q, K, V) = \text{softmax}(QKT\sqrt{d_k})V$$

Instead of performing only single attention in multi-head attention, the single attentions are performed in parallel and then concat.

$$h_i = Att(q_w q_i, k_w k_i, v_w v_i)$$

$$MHA(q, k, v) = \text{concat}(h_1, h_2, h_3, h_4)$$

The BERT has the feed-forward neural network. It has the transition layers with RELU as activation. Feed forward neural network of x

$$FFN(x) = (RELU(W_1 x + b_1))W_2 + b_2$$

$$= \max(0, xW_1 + b_1)W_2 + b_2$$

2) Machine learning Feature Extraction:
Term frequency–inverse document frequency (TF-IDF) is applied to extract machine learning model features. The TF-IDF converts each word into the form of vectors. The value for each word depends on the number of times the word is repeated in the data. Mathematically it is defined as [29]

$$TF(t, d) = \left(\frac{\text{The number of times the phrase t appears in document d}}{\text{Count of words in the document d}}\right)$$

$$IDF(t) = \log_e\left(\frac{\text{The amount of documents in the corpus as a whole}}{\text{The number of papers containing the phrase t}}\right)$$

III. Experimental Results and Discussions

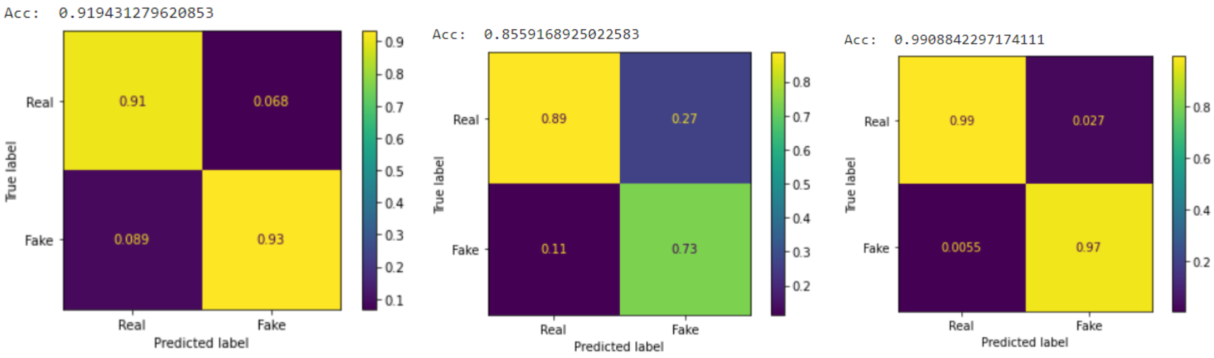
This section investigates four deep learning models (i.e., LSTM, Bi-LSTM, CNN Bi-LSTM, and BERT) to detect fake news. Moreover, we also examine the performance of the well-known machine learning models on unseen test data to provide whether the trained model could perform satisfactorily on real-world news.

Experimental Setup: The machine learning and deep learning models are trained with 80% of the data. Furthermore, 10% of data is used for validation and tested on 10% of the unseen data. The machine learning models are trained using the scikit-learn library in Python 3, and all deep learning models are trained in Tensorflow and PyTorch libraries.

A. Experiment I: PolitiFact Dataset

Table I presents the accuracy, precision, recall, and F1 score by training on the PolitiFact dataset for different machine and deep learning algorithms. It is noticed that the fine-tuning of BERT performs better with an accuracy of 91.94%, precision of 93.05%, recall of 91.0 9%, and F1 score of 92.06%. The other three deep learning algorithms, namely LSTM, Bi-LSTM, and CNN Bi-LSTM, provide results ranging the accuracy from 81.51% to 84.36%, precision from 81% to 84%, recall from 82% to 84%, F1 score from 81% to 84%. Simulation results indicate that BERT has 7% more accuracy than other deep learning models.

The fine-tuning of BERT achieves the best result since it can take bi-directional inputs simultaneously. It is pre-trained



(a) PolitiFact (b) GossipCop (c) CoAID dataset

Fig. 7: Confusion matrices simulated on the unseen test data by BERT algorithm. The figure represents the confusion matrices of the BERT in all three experiments.

TABLE I: Results on PolitiFact test Dataset

Algorithms	Accuracy	Precision	Recall	F1-Score
PAC	84.84%	85%	85%	85%
LR	81.06%	81%	81%	81%
NB	85%	85%	85%	85%
DT	76.89%	77%	77%	76%
GB	75.37%	75%	75%	74%
RF	78.40%	79%	78%	77%
KN	76.89%	77%	77%	76%
SVC	83.33%	83%	83%	83%
AB	75.75%	76%	76%	76%
XGBOOST	75.75%	75%	76%	75%
LSTM*	84.36%	84%	84%	84%
Bi-LSTM*	82.93%	83%	83%	83%
CNN Bi-LSTM*	81.51%	81%	82%	81%
BERT	**91.94%**	**93.05%**	**91.09%**	**92.06%**

* indicates the algorithm is performed with Gensim vectors

TABLE II: Results on CoAID Test Dataset

Algorithms	Accuracy	Precision	Recall	F1-Score
PAC	94.24%	94%	94%	94%
LR	90.45%	91%	90%	89%
NB	85%	88%	85%	81%
DT	90.09%	90%	90%	89%
GB	90.52%	91%	91%	89%
RF	90.52%	91%	91%	89%
KN	90.16%	90%	90%	89%
SVC	93.73%	94%	94%	93%
AB	91.62%	91%	92%	91%
XGBOOST	90.74%	90%	91%	90%
LSTM*	95.73%	95%	96%	95%
Bi-LSTM*	94.71%	95%	95%	94%
CNN Bi-LSTM*	93.35%	93%	93%	93%
BERT	**99.09%**	**97.35%**	**99.45%**	**98.38%**

* indicates the algorithm is performed with Gensim vectors

on natural language, whereas other deep learning and machine learning models are not pre-trained; thus, it performs better. Furthermore, BERT has self-attention, and no locality bias, which gives equal importance to the words in shorter and longer distances. On the other hand, LSTM is uni-directional, not a pre-trained technique. Moreover, Bi-LSTM and CNN Bi-LSTM are bi-directional but have the absence of attention and are not pre-trained in the natural language. The self-attention mechanism is absent in both techniques and couldn't offer equal importance for shorter and longer distances.

Besides, out of all the ten machine learning models, PAC, LR, NB, and SVC provide better accuracy compared to others ranging from 81.06% to 84.84%. It can be noted that the performance of the PAC is better among all the machine learning models, with an accuracy of 84.84%. Table I shows that other machine learning algorithms are found to be lower than 80% of accuracy. The GB, AB, and XGBOOST have low performance among all the machine learning models, with an accuracy of around 75%, i.e., only three by fourth predictions are correct.

B. Experimental II:(CoAID (Covid-19 Healthcare Misinformation Dataset))

Table II demonstrates the accuracy, precision, recall, and F1 score by training on the CoAID data set for various learning models. It is observed that the fine-tuning of the BERT gives the best results with an accuracy of 99.09%, precision of 97.35%, recall of 99.45%, and F1 score of 98.38%. The other three deep learning algorithms achieve an accuracy ranging from 93.35% to 95.53%, which reveals that BERT provides 5% better performance in terms of accuracy. The justification of these results is similar to the BERT characteristics (e.g., the capability of bi-directional, attention, pre-trained, etc.) described previously for the PolitiFact dataset in Table I. Table II also illustrates that machine learning models SVC, PAC, and NB deliver better accuracy results ranging from 91.62% to 94.24% than other machine learning models.

C. Experimental III: GossipCop

The simulated results by training on the GossipCop data set are presented in Table III for various learning algorithms. It is reported that identical to results obtained for PolitiFact, and CoAid datasets in Table I and Table II, respectively, BERT outperforms for GossipCop dataset. For instance, finetuning of BERT delivers an accuracy of 85.59%, a precision of 76.70%, a recall of 89%, and F1 score of 82.40%. The other deep learning models, LSTM, Bi-LSTM, and CNN Bi-LSTM performances range from 79.78% to 81.02%. For the machine algorithms, SVC executes the best result with an accuracy of

TABLE III: Results on GossipCop test data

Algorithms	Accuracy	Precision	Recall	F1-Score
PAC	79.53%	80%	80%	80%
LR	84.40%	84%	84%	83%
NB	82%	83%	82%	78%
DT	77.48%	77%	77%	77%
GB	81.75%	82%	82%	78%
RF	83.30%	82%	83%	82%
KN	78.78%	79%	79%	79%
SVC	84.98%	84%	85%	84%
AB	82.29%	81%	82%	81%
XGBOOST	81.68%	80%	82%	80%
LSTM*	79.78%	80%	82%	80%
Bi-LSTM*	78.56%	79%	79%	79%
CNN Bi-LSTM*	81.02%	81%	81%	81%
BERT	**85.59%**	**76.70%**	**89%**	**82.40%**

* indicates the algorithm is performed with Gensim vectors

84.98%, precision of 84%, recall of 85%, and F1 score of 84%.

Fig.7 (a) illustrates the confusion matrix for the BERT model with 91% true positives, i.e., it can predict 91% of the real news as accurate, 8.9% false negatives indicate 8.9% of trustworthy information as fake news. Similarly, 93% true negatives imply that it can envision 93% of fake news as fake news, and 6.8% false positives show the model may predict 6.8% of fake news as real news. Likewise, Fig.7(b) and Fig.7(c) depict the confusion matrix of the BERT model for the GossipCop and CoAID datasets, respectively.

Table IV summarizes the accuracy values of different learning strategies studied in literature to detect fake news. Table IV reveals that accuracy above 95% is reported by LSTM with Glove embeddings, Glove pre-trained embedding with FDNet, Distil BERT, and TFIDF with CNN. On the other hand, our experiments present an accuracy of 99.09% by fine-tuning BERT. It is interesting to note that the accuracy value of LSTM with Glove embeddings is very close to the proposed work since Glove embeddings are pre-trained on a huge number of Twitter datasets for social media-related applications. The other algorithms, i.e., SVM, FND NS, Random Forest, Recurrent neural networks, and Bi-LSTM, achieved results less than or equal to 95%. Compared to these models, the proposed model has a significant improvement of more than 4-5% of accuracy. Comparison with the same dataset: Table V compares accuracy with [17], which used the same dataset. The fine-tuning of BERT significantly improves the results on all three datasets compared to results reported by [17]. We record an improvement of 8.01%, 1.77%, and 0.5% on the datasets PolitiFact, GossipCop, and CoAID, respectively.

TABLE IV: Comparison with the result and those presented in the Literature

Model Name	Accuracy
Distil BERT [15]	97.20%
Glove pre trained embedding with FDNet neural network [16]	98.36%
Bi LSTM [21]	90.40%
Random Forest [19]	84.40%
Recurrent neural networks and transformer based models [20]	95% (f1-score)
TFIDF with CNN [18]	96.89%
FND NS [14]	74.8%
SVM [8]	94.19%
LSTM with Glove embeddings [17]	98.6%
Proposed method	**99.09%**

TABLE V: Comparison with the results of the same datasets in the literature

Name of the Dataset	LSTM with Glove Embeddings	Fine tuning of BERT (proposed method)	Improvement
Politifact	83.93% [17]	91.94%	**8.01%**
GossipCop	83.82% [17]	85.59%	**1.77%**
CoAID	98.6% [17]	99.09%	**0.5%**

IV. Conclusion

Currently, much fake news is uploaded on social media; consequently, it is challenging to check whether every piece of information is accurate. Hence, it will be helpful to develop automated fake news detection. In this context, we investigated different deep-learning algorithms to detect fake news on three real-world datasets: GossipCop, CoAID, and PolitiFact. In this work, all learning algorithms are trained and tested on unseen real-world datasets to estimate whether the models can perform in real-life conditions. We preprocessed the data, which includes everything from eliminating raw data to lemmatization for all three datasets. Machine learning models rely on the TF-IDF for feature analysis, whereas deep learning models depend on word embedding. The model's performance is measured in terms of accuracy, precision, recall, F1-score, and confusion matrix. Simulation results indicated that BERT performs better as compared to other learning techniques for all three datasets, i.e., GossipCop, CoAID, and PolitiFact.

References

[1] K. Rapoza. "forbes". https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stockmarket/#5e66ab9d2fac, 2017. [Online accessed 2017-09-06].

[2] Janna Anderson and Lee Rainie. "the future of truth and misinformation online". https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/, 2017. [Online accessed 2017-10-19].

[3] D. Russell. Benjamin franklin / the problem of 18th century fake news / ladybugs. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm, 2017. [Online accessed 2017-03-01].

[4] http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm.

[5] https://en.wikipedia.org/wiki/Fake_news.

[6] EM Okoro, BA Abara, AO Umagba, AA Ajonye, and ZS Isa. A hybrid approach to fake news detection on social media. *Nigerian Journal of Technology*, 37(2):454–462, 2018.

[7] Barbara Probierz, Piotr Stefański, and Jan Kozak. Rapid detection of fake news based on machine learning methods. *Procedia Computer Science*, 192:2893–2902, 2021.

[8] Pervaiz Akhtar, Arsalan Mujahid Ghouri, Haseeb Ur Rehman Khan, Mirza Amin ul Haq, Usama Awan, Nadia Zahoor, Zaheer Khan, and Aniqa Ashraf. Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research*, pages 1–25, 2022.

[9] M Sudhakar and KP Kaliyamurthie. Effective prediction of fake news using two machine learning algorithms. *Measurement: Sensors*, 24:100495, 2022.

[10] Rubita Sudirman, Narges Tabatabaey-Mashadi, and Ismail Ariffin. Aspects of a standardized automated system for screening children's handwriting. In *2011 First International Conference on Informatics and Computational Intelligence*, pages 49–54. IEEE, 2011.

[11] Saeed Amer Alameri and Masnizah Mohd. Comparison of fake news detection using machine learning and deep learning techniques. In *2021 3rd International Cyber Resilience Conference (CRC)*, pages 1–6. IEEE, 2021.

[12] Antonio Galli, Elio Masciari, Vincenzo Moscato, and Giancarlo Sperlí. A comprehensive benchmark for fake news detection. *Journal of Intelligent Information Systems*, 59(1):237–261, 2022.

[13] Shaina Raza and Chen Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362, 2022.

[14] Muhammad F Mridha, Ashfia Jannat Keya, Md Abdul Hamid, Muhammad Mostafa Monowar, and Md Saifur Rahman. A comprehensive review on fake news detection with deep learning. *IEEE Access*, 9:156151–156170, 2021.

[15] Jackie Ayoub, X Jessie Yang, and Feng Zhou. Combat covid-19 infodemic using explainable natural language processing models. *Information processing & management*, 58(4):102569, 2021.

[16] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44, 2020.

[17] Diaa Salama Abdelminaam, Fatma Helmy Ismail, Mohamed Taha, Ahmed Taha, Essam H Houssein, and Ayman Nabil. Coaid-deep: an optimized intelligent framework for automated detecting covid-19 misleading information on twitter. *IEEE Access*, 9:27840–27867, 2021.

[18] Lovedeep Singh. Fake news detection: A comparison between available deep learning techniques in vector space. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pages 1–4. IEEE, 2020.

[19] Muneer Bani Yassein, Shadi Aljawarneh, and Yarub Wahsheh. Hybrid real-time protection system for online social networks. *Foundations of science*, 25:1095–1124, 2020.

[20] Maha Al-Yahya, Hend Al-Khalifa, Heyam Al-Baity, Duaa AlSaeed, and Amr Essam. Arabic fake news detection: comparative study of neural networks and transformer-based approaches. *Complexity*, 2021:1–10, 2021.

[21] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[22] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*, 2020.

[23] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, 53(4), jul 2020.

[24] Abdullah Aziz Sharfuddin, Md Nafis Tihami, and Md Saiful Islam. A deep recurrent neural network with bilstm model for sentiment classification. In *2018 International conference on Bangla speech and language processing (ICBSLP)*, pages 1–4. IEEE, 2018.

[25] Jason P. C. Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns, 2015.

[26] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Miguel A Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348, 2021.