Knowledge-Based Systems 274 (2023) 110642

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Towards COVID-19 fake news detection using transformer-based models

Jawaher Alghamdi [a,b,*], Yuqing Lin [a], Suhuai Luo [a]

[a] School of Information and Physical Sciences, The University of Newcastle, Newcastle, Australia
[b] Department of Computer Science, King Khalid University, Abha, Saudi Arabia

ARTICLE INFO

Article history:
Received 10 November 2022
Received in revised form 17 April 2023
Accepted 13 May 2023
Available online 19 May 2023

Keywords:
COVID-19
Fake news
Misinformation
Pre-trained transformer models
Social media

ABSTRACT

The COVID-19 pandemic has resulted in a surge of fake news, creating public health risks. However, developing an effective way to detect such news is challenging, especially when published news involves mixing true and false information. Detecting COVID-19 fake news has become a critical task in the field of natural language processing (NLP). This paper explores the effectiveness of several machine learning algorithms and fine-tuning pre-trained transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT) and COVID-Twitter-BERT (CT-BERT), for COVID-19 fake news detection. We evaluate the performance of different downstream neural network structures, such as CNN and BiGRU layers, added on top of BERT and CT-BERT with frozen or unfrozen parameters. Our experiments on a real-world COVID-19 fake news dataset demonstrate that incorporating BiGRU on top of the CT-BERT model achieves outstanding performance, with a state-of-the-art F1 score of 98%. These results have significant implications for mitigating the spread of COVID-19 misinformation and highlight the potential of advanced machine learning models for fake news detection.

1. Introduction

COVID-19 was declared a Public Health Emergency of International Concern by the World Health Organization (WHO) on 30 January 2020 [1]. Due to the physical restrictions imposed by governments to reduce the impact of COVID-19, people tend to rely more on social media as the primary source of communication. A recent study found that average user activity on social media has increased by 25% due to the global lockdown [2]. Increasing concerns related to COVID-19 have prompted people to seek and share information about the pandemic on social media [3]. However, the dark side of the coin is fake tweets dissemination that spreads fear and panic about such a pandemic. The dissemination of such falsified tweets has led to various adverse consequences, including vaccination hesitancy [4], changes in health behaviour intentions [5], also some falsified information propagated about chloroquine's effectiveness has led to an increase of cases of chloroquine drug overdose [6]. In less than two months, the International Fact-Checking Network (IFCN) at the Poynter Institute found over 3500 false claims related to COVID-19 [7].

The coronavirus-related misinformation may have led to more than 800 deaths worldwide in the first three months of 2020.[1] Social media platforms differ from traditional news outlets in that they have a tendency to spread false information based on some characteristics more than the latter. Several recent WHO reports describe the spread of misinformation related to COVID-19 as an *Infodemic* which is defined as "an overabundance of information, both online and offline. It includes deliberate attempts to disseminate wrong information to undermine the public health response and advance alternative agendas of groups or individuals".[2]

The massive amount of user-generated content becomes prohibitively cumbersome to manually process due to the large volume of data. Therefore, it is imperative to develop automated fake news detection systems that detect fake content effectively.

However, detecting fake news on social media becomes even more challenging when considering the poor quality of user-generated content, complex semantics of natural language and the high dimensionality of textual data, especially when malicious entities can frequently manipulate and change their writing style to mimic trustworthy content [8].

As a key element of the detection approaches, researchers have proposed different ways to interpret the meaning of a word by representing it as an embedding vector. For learning word embeddings from large word corpora, neural network-based methods (e.g., Word2Vec [9]) and count-based models

* Corresponding author at: School of Information and Physical Sciences, The University of Newcastle, Newcastle, Australia.
E-mail addresses: jawaher.alghamdi@uon.edu.au (J. Alghamdi), yuqing.lin@newcastle.edu.au (Y. Lin), suhuai.luo@newcastle.edu.au (S. Luo).

[1] https://www.bbc.com/news/world-53755067.

[2] https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation.

https://doi.org/10.1016/j.knosys.2023.110642

(e.g., GloVe [10]) are commonly used. The downside of these embedding models is that they are context-free, i.e., context is neglected, and a static embedding for the words is generated regardless of their contexts. Thus, a model that can capture deep semantic and contextualised word embeddings is required for more fine-grained detection performance. Lately, the concept of *attention* has received more and more attention, and the natural language processing (NLP) community is starting to approach a paradigm shift, developing a set of models that not only improve accuracy but also address the problem of lacking labelled data, which has been a well-known problem in the NLP research.

Automatic detection of fake news is a non-trivial task, given that existing [deep] machine learning models (prior to the advent of almost ubiquitous transformer models) are impotent towards providing a deeper semantic understanding of input text. To respond to this, the NLP research has made great strides by introducing the transformer architecture. In addition, pre-trained language models (PLMs) that have been trained on massive unlabelled corpora are a current trend for text classification tasks. Such models have made a great breakthrough in many NLP tasks where the PLMs can be easily fine-tuned on many different NLP tasks. The main idea is to extract the pre-trained neural network layers from the language model (LM) and stack new neural network layers on top of them to adapt for the downstream task [11].

While deep learning (DL) approaches allow for capturing more salient and relevant information, transformer-based approaches have the power to encode deeper semantic and contextualised information about a given input text. To take advantage of the former and the latter, we propose different architectures using different neural-based structures on top of the PLMs. This paper explores the potential of various PLMs such as CT-BERT [12], RoBERTa [13], and BERT-based classifiers for detecting COVID-19 fake news. We introduce different downstream neural network structures on top of BERT [14] and CT-BERT (fine-tuning strategies) to examine how effective different downstream neural network structures, added on top of BERT and CT-BERT architectures, are in improving COVID-19 fake news detection.

Our experiments have demonstrated (see Section 3.5.3) that different pre-trained transformer-based models perform differently under different strategies. For example, using the pre-trained CT-BERT with a Bidirectional Gated Recurrent Units (BiGRU) layer on top of it has shown to work the best among all other proposed strategies and official baseline models. Generally speaking, extending the PLMs using more expressive powered architectures such as CNN and BiGRU shows promising results; of course, the potential issue of these architectures is the excessive number of parameters used, which slows down the training process. Also, we have demonstrated that fine-tuning the PLMs produce better results than feature-based approaches by a clear margin.

The main contributions of this work may be resumed as:

1. To obtain the best performance, we have explored novel transferring transformer-based methodologies based on different downstream neural network structures.
2. Exploring the effects of fine-tuning approach on the proposed transferring methodologies to support the stated hypothesis that fine-tuning approach has good potential for further improving the performance.
3. Extensive experiments presented using different classical and advanced machine learning models, and the performance differences between these widely used models and the proposed models are compared and analysed.

This paper is structured as follows. Section 2 reviews previous studies on the topic. Section 3 describes the methodology and covers the analysis of the results. Section 4 provides the discussion and insights. Finally, Section 5 concludes the paper.

2. Related work

This section presents a brief overview of the related work directly relevant to Constraint@AAAI2021 COVID19 [15], with a particular emphasis on transformer-based approaches that utilise this dataset. It also highlights a selection of prior work that uses classical and advanced machine learning-based fake news detection approaches. There are different approaches that have been employed to detect fake news, ranging from metaheuristic-based methods to traditional machine learning (ML) methods and the-state-of-the-art transformer-based models.

2.1. Fake news detection

2.1.1. Classical and advanced machine learning approaches

Metaheuristic-based methods have been utilised as an intriguing solution search strategy for detecting fake news in online social networks and media. For example, a study by [16] proposed a novel model for detecting fake news that uses optimisation methods. However, metaheuristic-based methods need more time and computational resources to explore the solution space compared to traditional ML approaches. Statistical ML methods such as Logistic Regression (LR), Support Vector Machines (SVMs), K-Nearest Neighbour (K-NN), Naive Bayes (NB), Decision Trees (DT), Random Forest (RF), Gradient Boost (GB), and XGBoost (XGB) have traditionally been used in text classification. The goal of the Constraint@AAAI2021 COVID19 fake news detection challenge [17] is to develop a model capable of distinguishing between real and fake news related to COVID-19. As part of an ensemble constructed in [18], the team used bidirectional Long Short-Term Memory (BiLSTM), SVMs, LR, NB, and NB combined with LR, achieving a 0.94 F1 score.

Fake COVID-19-related news was examined by [19], in which the authors obtained data from 150 users by extracting information from their social media accounts, such as Twitter, email, mobile, WhatsApp, and Facebook, for a period span from March 2020 to June 2020. They removed information irrelevant to the COVID-19 data and incomplete news during the pre-processing phase. The classification was performed using K-NN, where the results showed the best prediction scores for June with a 0.91 F1 score and the worst ones for March with a 0.79 F1 score. Compared in [15] are four ML baseline models, namely, DT, LR, GB and SVMs, to detect COVID-19-related fake news with the best performance of 93.46% F1-score with SVMs. In [20], the authors applied classical ML algorithms using several linguistic features, such as n-grams, readability, emotional tone and punctuation. Their experimental results found a linear SVM to be the best performing model with a weighted average F1 score of 95.19% on test data.

In recent years, researchers have increasingly focused on deep neural networks (DNN)-based models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and BiLSTM that combine multiple DNN configurations. The technique of learning how to transfer knowledge is a concept in ML known as *transfer learning*, which stores and applies the knowledge gained from performing a specific task to another problem. Learning this way is useful when it comes to training and evaluating models with relatively small amounts of data. In the area of NLP, *transfer learning* is achieved by creating a set of pre-trained embedding models trained on a massive amount of text. Several neural-based models have been proposed to model the COVID-19 fake news detection problem using a context-independent pre-trained word embedding layer [15,18,20]. As these basic models are context-free, they are impotent towards capturing deep contextualised trends in the input text. Therefore, researchers developed attention-based models

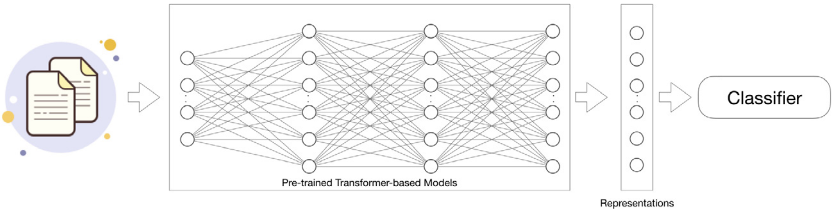


Fig. 1. PLMs fine-tuning.

that can provide context-aware word embeddings pre-trained on large-scale datasets, which are paramount for the success of most NLP tasks. The main objective of this paper is to investigate the use of advanced ML models and transfer learning in evaluating the credibility of news content related to COVID-19.

2.1.2. Transfer learning

A text mining model is a method for extracting useful information and knowledge from unstructured text [21]. Text mining models have become much more sophisticated with the advancements in DL techniques used in NLP. As of recently, with the advent of transformer-based structures, PLMs have become mainstream for downstream text classification [14]. For example, major advances have been driven by the use of PLMs, such as ELMO [22], GPT [23] or BERT. BERT and RoBERTa, as the most commonly utilised PLMs, were trained on exceptionally large corpora. The success of such approaches raises the question of how to use such models for downstream text classification tasks? Over the PLMs, task-specific layers are added for each downstream task, and then the new model is trained with only those layers from scratch [12–14] in a supervised manner; see Fig. 1.

More specifically, these models exhibit a two-step learning approach. They learn pre-trained language representations by analysing much text in a self-supervised fashion. This process is commonly called pre-training. Then these pre-trained language representations can be applied to downstream NLP tasks by selecting either of two approaches: feature-based and fine-tuning. The former uses pre-trained representations and includes them as additional features for learning a given task. The latter introduces minimal task-specific parameters, and all pre-trained parameters are fine-tuned on the downstream tasks. The advantage of such transfer learning is that the deep context-aware word representations can be learned from large unannotated text corpora in self-supervised pre-training; this is especially useful when learning a domain-specific language with insufficient labelled data.

Besides the fact that surface-level features cannot effectively capture semantical patterns in text, the lack of a sufficient amount of data constitutes a bottleneck for the advanced ML models. Thus, to address this, we exploit the power of BERT and its variations in building robust fake news predictive models. Relatively little research has been done to detect fake news using the recent pre-trained transformer-based models. The few observational studies that have been done using such models, despite the use of different methodologies and different scenarios, have shown promising results.

One recent example of this is a study conducted by Aggarwal et al. [24] showed that BERT, even with minimal text pre-processing, provided better performance compared to that of LSTM and gradient boosted tree models. Similarly, Jwa et al. [25] achieved a high F1 score of 0.746 for fake news detection using BERT on the FNC dataset by analysing the relationship between the headline and body text of news. Baruah et al. [26] also utilised BERT for the classification task of automatically detecting fake news spreaders, achieving an accuracy of 0.690. However, BERT is computationally expensive due to its millions of parameters (e.g., BERT$_{BASE}$ has 110 million parameters while BERT$_{LARGE}$ has 340 million parameters) [14]. DistilBERT [27], a variation of BERT, reduces its size by 40% while retaining 97% of its language understanding abilities, resulting in faster training (60% faster). Another robust BERT model, RoBERTa [13], was developed using a larger dataset, larger batches, and more iterations.

In [28], the authors applied a pre-trained transformer model, so-called XLNet, combined with Latent Dirichlet Allocation (LDA) by integrating contextualised representations generated from the former with topical distributions produced by the latter. Their model achieved an F1 score of 0.967 on the Constraint@AAAI2021-COVID19 fake news dataset. Using a combination of existing contextual representations, such as BERT, and knowledge graph-based representations, a study by [29] achieved an accuracy of 95.70% and an F1 score of 95.69% on the same dataset. In the same vine, a fine-tuned transformer-based ensemble model has been proposed by [30]. The proposed model achieved an F1 score of 97.9% using the same dataset. The ensemble transformer model differs from the previously mentioned models in that it can combine the advantages of multiple transformer models, leading to an enhanced overall performance. This could explain its achievement of state-of-the-art performance. A compact overview of the ML models proposed by related work is shown in Table 1.

To this end, classical ML algorithms are easy to comprehend and perform well on small datasets, but they (i) require complex feature engineering and (ii) fail to capture substantial semantical contextual knowledge for a specific input text. To overcome this, advanced ML techniques such as CNNs and RNN-based methods are well suited for complicated classification problems, powered by a massive amount of data, and can learn more complicated (latent) features. However, even though CNNs have proven effective in extracting local features, they typically struggle with capturing long-term contextual dependencies. In contrast, RNN-based methods perform sub-optimally in handling such dependencies and are not a good candidate for capturing local features. As such, a combination of these two architectures may be able to overcome some of their inherent limitations. Plus, adding this unified architecture on top of a transformer-based model such as BERT (more specifically, the variant trained on massive COVID-19 tweets) would give the model far more expressive power by allowing it to learn deep and meaningful insights for a given input text. As stated in the introduction and different from these studies mentioned above, we aim to exploit the power of the advanced ML models in combination with the deep contextualised PLMs by testing the effectiveness of different downstream neural network architectures added on top of different transformer-based models for COVID-19 fake news detection. A comparison of the applied approach with the state-of-the-art on COVID-19 dataset is shown in Table 2.

3. Methodology

3.1. Problem statement

Suppose we have a set $T = \{t_1, t_2, .., t_L\}$ of L labelled news (i.e., tweets) either fake or real, where L is the total number of

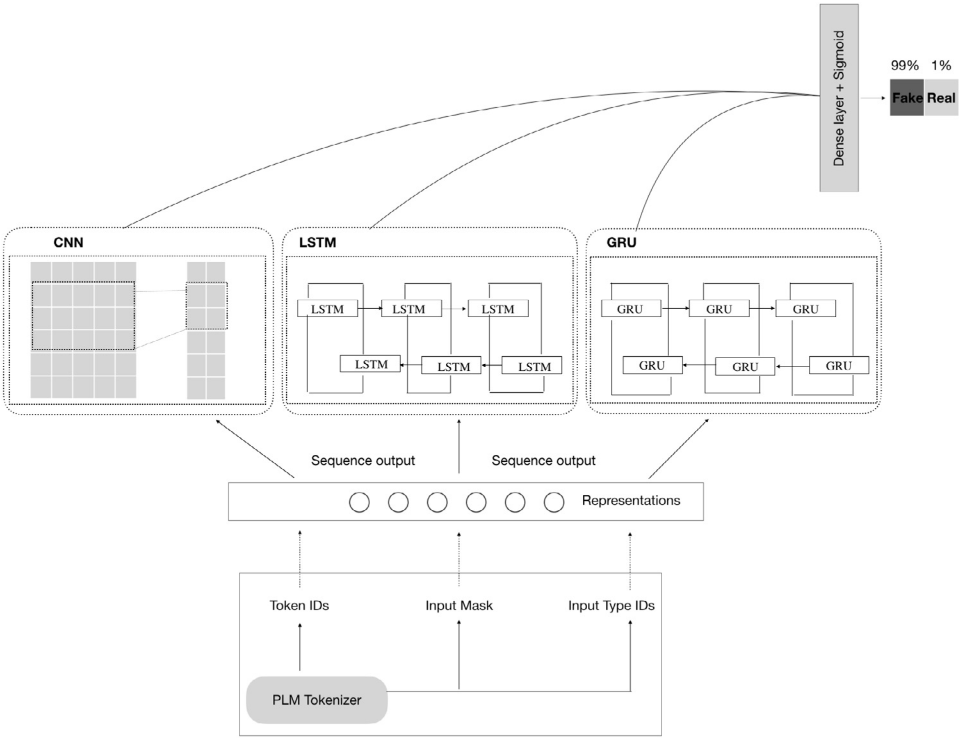


Fig. 2. PLMs + downstream neural network structures.

Table 1
A compact overview of ML models applied by the related work.

Ref	Model	Advantages	Disadvantages
[18]	Ensemble LR+SVMs+NB+BiLSTM	The ensemble method combines the strengths of multiple models and helps to reduce the impact of individual weaknesses and can improve the overall accuracy of the model.	Ensemble may be more computationally intensive and require more resources than a single model.
[19]	KNN	KNN is a simple and intuitive algorithm that is easy to understand and implement.	KNN is a lazy learner, which means it requires a significant amount of time to make predictions compared to other algorithms.
[15,20]	SVM	SVMs can effectively model complex decision boundaries and have been shown to perform well on high-dimensional datasets.	SVMs can be computationally expensive, especially on large datasets.
[29]	BERT	BERT can handle complex language structures, including context-dependent meanings of words, idiomatic expressions, and long-range dependencies.	BERT requires a large amount of computational resources for training and inference, making it difficult to deploy on resource-constrained systems.
[28]	XLNet	XLNet can capture bidirectional context more effectively than BERT, allowing it to better handle tasks that require understanding of long-term dependencies.	XLNet requires significant computational resources and may not be suitable for all applications.
[30]	Transformer ensemble	Transformer ensemble models can combine the strengths of multiple pre-trained transformer models to achieve better performance on downstream tasks.	Ensemble models can be computationally expensive and require large amounts of memory, making them difficult to deploy on resource-constrained devices.

Table 2
A comparison of the best performing applied approach with the state-of-the-art on COVID-19 dataset.

Ref.	Model	F1 score (%)	Features used	Adding downstream neural based structures on top of transformers
[15]	SVMs	93.46%	TF-IDF	N/A
[18]	Ensemble LR+SVMs+NB+BiLSTM	94%	TF-IDF	N/A
[20]	SVMs	95.19%	N-grams, readability, emotional tone and punctuation	N/A
[29]	BERT	95.69%	Context-aware and knowledge graph representations	No
[28]	XLNet	95.70%	Context-aware representations and topical features	No
[30]	Transformer ensemble	97.9%	Context-aware representations	No
Ours [5]	**CT-BERT+BiGRU**	**98.54%**	**Context-aware representations**	**Yes**

news. The task is to learn fake news detection function $f(T) \rightarrow \hat{y}$, such that it maximises prediction accuracy. Here, we cast the problem as a binary classification where a piece of information could be fake ($\hat{y} = 0$) or real ($\hat{y} = 1$).

We investigate and compare the effectiveness of a set of different PLMs using different (simple and sophisticated) downstream neural network structures. In particular, we extend the pre-trained BERT and CT-BERT LMs by adding various downstream neural network architectures (See Fig. 2). We test the effectiveness of such architectures using both frozen (feature-based approach) and unfrozen weights (fine-tuning approach where all model parameters are jointly trained on a given supervised task).

Using BERT coupled with various advanced ML models, we fine-tune these variants to predict whether the information is fake or not by adding a classification (output) layer on top of each variant with a sigmoid function. The outcome of the activation of the output layer is what will be shown as the model's prediction. The workflow of the proposed models can be seen in Algorithm 1.

Algorithm 1 PLMs-based Downstream Neural Network Models

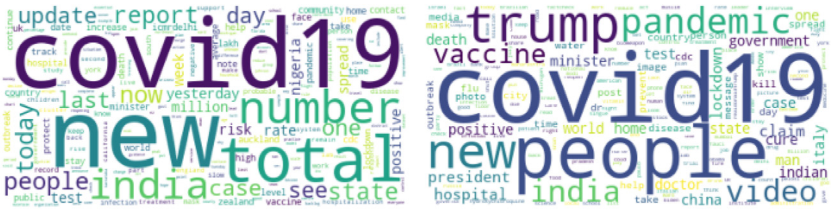
```
Data: COVID-19 dataset $C_D$: Training set $T_{train}$, Validation set $V_{valid}$ and Testing set $T_{test}$
Output: Probability $\mathbf{P} = \{p_i \in [0, 1]$; Truthfulness class (fake, real)
for $C \in$ preprocessed $C_D$ do
    Apply the corresponding model's tokeniser, generate token ids, attention mask and token type ids
    Generate word embedding vectors $\vec{w}_i$
end for
for $C \in T_{train}$ do
    Add downstream neural network architectures on top of PLMs    ▷ E.g., BiGRU, CNN etc.
    Add a dense layer
    Calculate the probabilities of labels using Sigmoid
end for
for $C \in T_{test}$ do
    Apply the trained model to predict the label (fake or real)
end for
```

3.2. The proposed models

- **BERT$_{BASE}$** [14]**:** This model, developed by Google AI, has proven to be a powerful tool for text classification [31, 32]. BERT is a multi-layer bidirectional transformer encoder trained on English Wikipedia and Book Corpus containing 2500M and 800M tokens. BERT$_{BASE}$ uses the transformer's encoder with 12 layers, 12 self-attention heads, and 110 million parameters. In our work, we used the uncased version of BERT, which is considered as a baseline model for BERT$_{BASE}$ model set. Input sequences of a maximum length of 128 are fed into BERT, and based on some analysis, 768-d vector representation is produced for each token. These vectors carry meaningful information about the context of each token. This model uses the corresponding [CLS] token's representation (a vector of size 768 representing the entire sequence) as input to the output layer. We also experiment with BERT$_{LARGE}$, which has 24 layers, 16 attention heads, and 340 million parameters.
- **BERT$_{BASE}$+CNN:** In this architecture, we extend the BERT model by adding a CNN layer for fine-tuning. First, the representations of the last transformer encoder (sequence output) are used as input to a Conv1D with 128 filters, each with a kernel size of 5, activated with the ReLU function. This is followed by a max-pooling layer to reduce the feature maps. Finally, the resultant feature map is flattened, and the output is passed to the output layer.
- **BERT$_{BASE}$+(Bi)LSTM:** Similar to the above architecture, all representations of the latest transformer encoder are used as input to a single (bidirectional) LSTM layer with 128 units, followed by an output layer.
- **BERT$_{BASE}$+(Bi)GRU:** The sequence output generated from the latest transformer encoder is used as an input to a (bidirectional) GRU layer with 128 units. The resultant hidden state is then fed into an output layer.
- **BERT$_{BASE}$+CNN-BiLSTM:** The sequence output generated from the latest transformer encoder is used as an input to a hybrid model consisting of a single CNN layer followed by a max-pooling layer. The output is then encoded using a BiLSTM layer.
- **BERT$_{BASE}$+CNN-BiGRU:** Similar to the previous model, with BiLSTM layer was replaced with a BiGRU one.
- **BERT$_{BASE}$+mCNN:** This model is defined with three input channels for processing different n-grams of the input text. Each of these channels consists of three layers: convolution that extracts different word n-gram features and a kernel size set to 4, 6, 8 g to read at once; a max-pooling layer to enable the extraction of the most salient features from each feature map; and a flatten layer to reduce the three-dimensional output into a two-dimensional one. Then the extracted features from the three channels are concatenated into a single vector, and the output is passed to a classification layer.
- **BERT$_{BASE}$+mCNN-BiLSTM:** We extend the previous model by passing the resulted single vector of the mCNN model to a BiLSTM layer. Apart from that, the configuration is identical to BERT$_{BASE}$+mCNN.
- **RoBERTa** [13]**:** Stands for the Robustly optimised BERT approach introduced by Facebook. It is simply retraining of BERT with improved training methodology (i) by removing the Next Sentence Prediction task from the pre-training process, (ii) RoBERTa was trained over ten times more data, and (iii) by introducing dynamic masking using larger batch sizes so that the masked token changes during the training rather than static masking pattern used in BERT. Thus, RoBERTa introduces a different pre-training approach to BERT. We experiment with the two variations: RoBERTa$_{BASE}$ and RoBERTa$_{LARGE}$.
- **CT-BERT** [12]**:** Stands for COVIDTwitter-BERT is a transformer-based model, pre-trained on a large corpus of Twitter posts (160M tweets) on the topic of COVID-19 collected from January 12 to April 16, 2020. Thus, the pre-trained CT-BERT has the same domain as the COVID-19 dataset used in this work; thus, we expect CT-BERT to provide better results than the general pre-trained BERT model. We extended the model with the same downstream neural structures used on top of BERT.

3.3. Experimental setup

All the work for the experiments was carried out using Intel Core i5 2.3 GHz, 8 GB RAM system running macOS. We implemented the transformer-based models using Tensorflow Hub—Tensorflow official model repository. Scikit-learn is used to implement classical ML algorithms, while Tensorflow and Keras libraries are used to implement the advanced ML models. This study used the common performance metrics to evaluate the performance of the proposed models, namely, accuracy, precision, recall, and F1 score.



(a)

(b)

Fig. 3. Wordcloud of (a) real news and (b) fake news.

Table 3
The statistics of COVID-19 dataset.

# Candidate news	6420
# True news	3360
# Fake news	3060

3.4. Dataset

A collection of COVID-19-related social media posts, comments and news, classified as real or fake, based on their truthfulness. The dataset [15] is collected from various social media platforms, such as Twitter and YouTube. The challenge organisers collected 10,700 social media posts and news articles about COVID-19 in the form of an annotated dataset in English. As the dataset has been separated in advance by the task organisers into training, validation, and testing sets, we opted to evaluate our models using the original split.

Fig. 3 depicted the corpus's word cloud representation of real and fake news, respectively. From the figure, it can be deduced that the most frequent words apart from "covid19" are "people", "India", "pandemic", "vaccine", "risk", "hospital", "government" and so on. The statistics of the dataset are shown in Table 3.

3.4.1. Preprocessing

Preprocessing of input text includes tokenising given text using the model's tokeniser to generate input ids, input masks, and input type ids for further processing. By design, BERT (and its variations) take a sequence of tokens with a maximum length of 512 and produce a representation of the sequence in a 768-dimensional vector for BERT$_{BASE}$ and 1024-dimensional vector for BERT$_{LARGE}$. Thus, the text must be padded or truncated to ensure that all sequences have similar lengths. In this case, all sequences have been truncated to a length of 128.[3]

Since user-generated content is often noisy and ambiguous, preprocessing the data is important before feeding it to the models. While removing the emojis/emoticons is common based on the assumption that it reduces the noise in data, this assumption does not always hold. As users tend to use emojis to express their emotions, the emojis may provide deep insights into a text (sentiment words), such as sentiment and emotions; thus, it is considered beneficial to retain them in such a way by converting them into text – using the Python library emoji – in our work.

The models are trained with a batch size of 4 for 3 epochs as we found it to be best for all models based on trial and error and as part of future experiments, we will investigate the selection of hyperparameter values. The sigmoid function is used in the output layer to reduce the error during training, while binary cross-entropy is used to calculate the loss during back-propagation. In our final configuration, we use Adam optimiser with a learning rate of $2 \times 10^{-}5$ for BERT and its variations while $1 \times 10^{-}5$ for training CT-BERT and RoBERTa models. Again, based on a trial-and-error examination, these values performed the best. Furthermore, these models were trained with minimal preprocessing since transformer-based models come with their own tokenisers and can handle punctuation and lowercase text.

3.5. Results

3.5.1. Classical and advanced ML approaches

We have conducted extensive experiments using classical and advanced ML models on the COVID-19 dataset. The former includes LR, SVMs, NB, RF, and XGB with two types of feature extractor methods, namely, CV and TF-IDF. The latter use GloVe with a 100-d vector to encode input text, which includes CNN, BiLSTM, BiGRU and a hybrid model of CNN with BiLSTM and BiGRU. Table 4 displays the summary of the results obtained. The table shows that the recurrent-based models [BiLSTM and BiGRU] reported significantly higher detection scores than other algorithms. This is followed by CNN and LR classifiers with an F1 score of 0.9463 and 0.9424, respectively. Based on the analysis, statistical methods such as LR and SVMs yield better results than the bagging and boosting techniques such as RF and XGB. Moreover, the findings demonstrate that using CV as a feature extractor is better than TF-IDF for the classical ML methods and fine-tuning the embedding layer during training is better than the static approach for the advanced ML models. We can see how much dynamic embedding layer contributes to the overall performance. Although we applied the same maximum length limit as in transformer-based models (i.e., 128), we can observe the power of the advanced ML models in capturing useful patterns.

It is well known that the size of the data has an impact on the performance of such approaches. Traditional ML algorithms, on the other hand, are less affected by the size of data. As such, advanced ML models typically outperform other approaches when there is a large amount of data. This can be attributed to the outstanding capability of such algorithms in automatically extracting a wide range of informative features. According to [33], supervised DL models will usually perform well at about 5000 examples per class, so such models may not be suitable for scenarios involving few labelled examples. Interestingly, other studies, such as [34], have shown that when one has only 100–1000 labelled examples per class, BERT is a more effective technique for document classification than other traditional ML approaches, owing to the benefits of pre-training and fine-tuning. Transfer learning is a technique that allows us to take advantage of the benefits of DL while using fewer amounts of data.

3.5.2. Transformer-based models

In this section, by comparing our proposed models to other official baselines and other counterpart algorithms, we try to answer the following questions through the experiments:

1. **EQ1.** Can CT-BERT trained on a corpus of COVID-19 tweets achieve high detection performance compared to the other PLMs trained on a generic language?

[3] We found that simple truncation worked well where we consider only the first 128 tokens while ignoring the rest.

Table 4
Performance comparison (%) of different classical and advanced ML models on the COVID-19 dataset.

Type	Method	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
Baselines ML (CV)	LR	0.9402	0.9348	0.9501	0.9424
	SVMs	0.9313	0.9304	0.9379	0.9341
	NB	0.9248	0.9420	0.9166	0.9291
	RF	0.9178	0.9143	0.9275	0.9209
	XGB	0.8874	0.8920	0.8928	0.8924
Baselines ML (TFIDF)	LR	0.9276	0.9437	0.9199	0.9317
	SVMs	0.9393	0.9473	0.9373	0.9423
	NB	0.9037	0.9652	0.8662	0.9130
	RF	0.9187	0.9250	0.9201	0.9225
	XGB	0.8841	0.8946	0.8852	0.8899
Baselines ML (GloVe, static)	CNN	0.8766	0.9100	0.8482	0.8780
	BiLSTM	0.9313	0.9140	0.9589	0.9359
	BiGRU	0.9294	0.9001	0.9732	0.9352
	CNN-BiLSTM	0.8715	0.8132	0.9795	0.8886
	CNN-BiGRU	0.9107	0.8894	0.9473	0.9174
Baselines ML (GloVe, dynamic)	CNN	0.9449	0.9647	0.9286	0.9463
	BiLSTM	0.9514	0.9504	0.9571	0.9537
	BiGRU	**0.9523**	**0.9465**	**0.9634**	**0.9549**
	CNN-BiLSTM	0.9393	0.9396	0.9446	0.9421
	CNN-BiGRU	0.9332	0.9469	0.9241	0.9354

Table 5
Performance comparison (%) of different PLMs with different downstream neural network architectures.

Models	Metrics			
	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
SVMs+LR+NB+biLSTM [18]	N/A	N/A	N/A	0.94
SVMs [20]	0.9570	0.9571	0.9570	0.9570
SNN(LM+KG) [29]	0.9570	0.9533	0.9652	0.9569
XLNet+LDA [28]	N/A	0.968	0.967	0.967
SVMs [15]	0.9332	0.9333	0.9332	0.9332
Ensemble transformers [30]	0.9799	0.9799	0.9799	0.9799
BERT$_{BASE}$	0.9617	0.9474	0.9812	0.9640
BERT$_{LARGE}$	0.9673	0.9589	0.9795	0.9691
BERT$_{BASE}$+CNN	0.9743	0.9776	0.9732	0.9754
BERT$_{BASE}$+LSTM	0.9710	0.9673	0.9777	0.9725
BERT$_{BASE}$+BiLSTM	0.9780	0.9752	0.9830	0.9791
BERT$_{BASE}$+CNN-BiLSTM	0.9771	0.9735	0.9830	0.9782
BERT$_{BASE}$+GRU	0.9654	0.9548	0.9804	0.9674
BERT$_{BASE}$+BiGRU	0.9808	0.9737	0.9902	0.9819
BERT$_{BASE}$+CNN-BiGRU	0.9664	0.9448	0.9938	0.9687
BERT$_{BASE}$+mCNN	0.9678	0.9574	0.9821	0.9696
BERT$_{BASE}$+mCNN-BiLSTM	0.9729	0.9601	0.9893	0.9745
DistilBERT	0.9617	0.9617	0.9652	0.9635
CT-BERT	0.9757	0.9692	0.9848	0.9770
CT-BERT+CNN	0.9822	0.9830	0.9830	0.9830
CT-BERT+LSTM	0.9762	0.9785	0.9759	0.9772
CT-BERT+GRU	0.9762	0.9652	0.9902	0.9775
CT-BERT+BiLSTM	0.9724	0.9538	0.9955	0.9742
CT-BERT+BiGRU	**0.9846**	**0.9797**	**0.9911**	**0.9854**
CT-BERT+CNN-BiLSTM	0.9645	0.9394	0.9964	0.9671
CT-BERT+CNN-BiGRU	0.9804	0.9695	0.9938	0.9815
CT-BERT+mCNN	0.9799	0.9795	0.9821	0.9808
CT-BERT+mCNN-BiLSTM	0.9799	0.9804	0.9812	0.9808
RoBERTa$_{BASE}$	0.9668	0.9541	0.9839	0.9688
RoBERTa$_{LARGE}$	0.9565	0.9581	0.9589	0.9585

2. **EQ2.** Do the sophisticated downstream neural network structures, compared to the superficial output dense layer, improve the model's performance?
3. **EQ3.** Does the fine-tuning approach improve the model's performance?

3.5.3. Fake news detection performance

Since CT-BERT has the same domain as the used COVID-19 dataset, we expect it to provide better results than other general PLMs. To answer **EQ1**, we first compare the proposed architectures with the official baselines. The effectiveness of different downstream neural network approaches is evaluated. The results are depicted in Table 5, with the best performance scores highlighted in bold. Based on the results, we have the following observations:

- The experimental results show that CT-BERT+BiGRU outperforms all models, which is not so surprising, given that CT-BERT, again, was pre-trained on a large corpus of Twitter posts on the topic of COVID-19. This seems to confirm our hypothesis that a model based on CT-BERT would perform better than other transformer-based approaches. However, a natural question arises: is this finding simply relate to the task of pre-training? Or there are other factors at play? Answering this question will form the basis of our future work.
- Moreover, GRU, as an improved version of LSTM, with bidirectionality, is capable of capturing informative features, leading to better detection performance. As such, it can be clearly seen from Table 5 that CT-BERT coupled with BiGRU achieved the state-of-the-art results, outperforming the official baselines and other algorithms.
- Additionally, to answer the question **EQ2**, we found that complex downstream neural structures added on top of PLMs perform better than simply adding a single output dense layer, i.e., CT-BERT + BiGRU > CT-BERT. This also applies to the BERT model where BERT$_{BASE}$ + BiGRU > BERT$_{BASE}$. This indicates that extending PLMs with more complex structures has the potential to capture more deep contextual information. This, of course, complicates the model conceptually and computationally, yet adding downstream neural structures is shown to be effective.
- Moreover, it seems that the base versions have better accuracy than their larger counterparts, which indicates that adding more model parameters increases the risk of performance degradation. However, this assumption does not always hold since (some) models with a smaller number of parameter values may produce lower accuracy compared to their larger counterparts. For example, we found recurrent neural-based models to outperform CNN-based models, see Table 4. Research [22] found that RNN is able to learn contextual representations in such a way by inspecting representations at different hidden layers.
- Furthermore, the results showed that BiGRU holds potential in learning CT-BERT features better than other algorithms, despite the little difference in their resulted F1 scores.
- Comparing LSTM and GRU, at this level of analysis, it seems there is no clear evidence that one consistently overperforms the other. However, GRU is shown to perform well and be more efficient given a small amount of training data. Furthermore, using bidirectionality shows outstanding performance since it captures information both from left and right, leading to learning more useful contextualised representations.
- Comparing the classical and advanced ML approaches with transformer-based methods, the latter outperforms the former with a clear margin. Indeed, in the case of Twitter (user-generated content often contains misspellings, noise, and abbreviations), one of the main advantages of BERT (and its variations) is the use of sub-tokens instead of a fixed per-word token; it is, thus, ideal for use with such a data [35] compared to the off-the-shelf context-independent word embeddings.

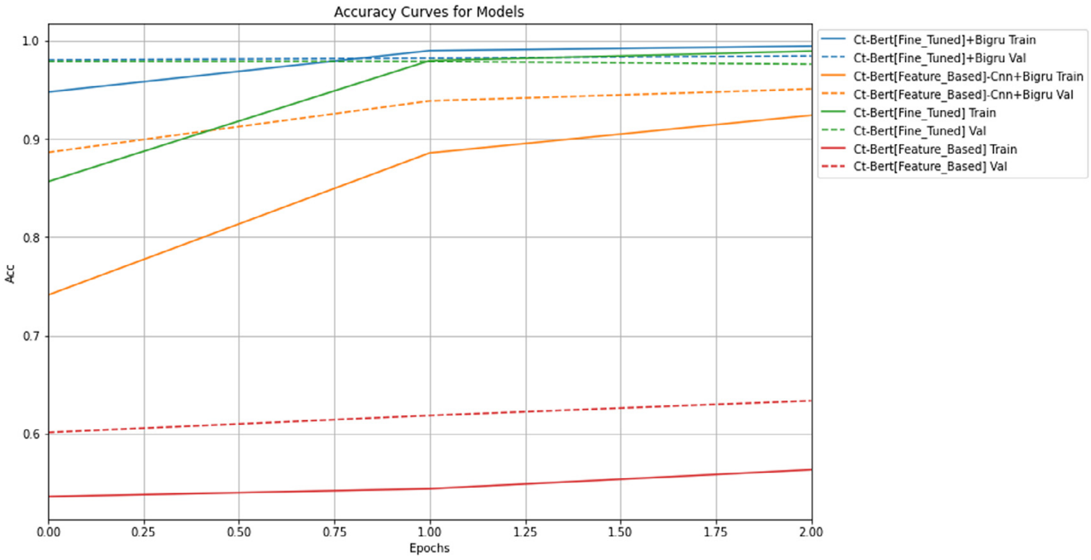


Fig. 4. Comparison (%) performance using *Accuracy* curves for models.

3.5.4. *Fine-tuning or not*

To answer **EQ3**, here, we study how fine-tuning impacts performance; we experimented with two approaches (i) we evaluate the role of the PLMs when they are frozen during the training phase (i.e., feature extractor) and (ii) we assess the impact of fine-tuning (unfreezing all parameters). As a feature extractor, all the layers of the pre-trained transformer-based model are frozen during the fine-tuning phase, and extra downstream neural layers, including an output sigmoid layer, can be added on top of the model and trained from scratch. That is, only the weights of the added layers will be updated during model training.

In the second approach, the entire pre-trained model has trained on our domain-specific dataset, and the output is fed into downstream neural layers, followed by a sigmoid layer. The error is then backpropagated throughout the architecture, and the model's pre-trained weights are updated based on the new dataset. In other words, the fine-tuning phase begins with the model parameters obtained from the pre-training, and all of the parameters are fine-tuned during that phase. To prove the robustness of the fine-tuning approach, we assess the impact of the two approaches, i.e. fine-tuning or not, using (some of) the proposed methodologies. Table 6 illustrates the comparative results between these two approaches. The experimental results clearly show the effectiveness of the fine-tuning approach. We further assess the influence of using downstream neural structures with and without fine-tuning. The comparative analysis is shown in Figs. 4 and 5. With the feature extractor approach, it is found that extending the pre-trained transformer-based models with sophisticated neural network structures provides better accuracy than extending such models with a simple output layer. To illustrate, as can be seen in Figs. 4 and 5, training the CNN-BiGRU model with CT-BERT as a feature extractor model yields (almost) equivalent results with slightly small differences in accuracy when compared to the fine-tuned CT-BERT+BiGRU model. This further amplifies the previous observation that downstream neural network structures are able to extract useful information effectively. In addition, fine-tuning the CT-BERT+output layer provides better performance than it does with the non-fine-tuning approach. Thus, we suspect fine-tuning is more critical as fine-tuned versions of DL outperform the frozen ones by a considerable margin.

Table 6
The results of the three best performing models with and w/o fine-tuning.

Models	Metrics			
	Acc.	Prec.	Rec.	F1
CT-BERT[a]	0.6364	0.6311	0.7348	0.6790
CT-BERT	0.9757	0.9692	0.9848	0.9770
BERT_{BASE}+BiLSTM[a]	0.9332	0.9208	0.9545	0.9373
BERT_{BASE}+BiLSTM	0.9780	0.9752	0.9830	0.9791
BERT_{BASE}+BiGRU[a]	0.9285	0.9290	0.9348	0.9319
BERT_{BASE}+BiGRU	0.9808	0.9737	0.9902	0.9819
CT-BERT+BiGRU[a]	0.9435	0.9222	0.9741	0.9475
CT-BERT+BiGRU	**0.9846**	**0.9797**	**0.9911**	**0.9854**

[a] Indicates models with frozen weights.

4. Discussion and insights

The study investigated the effectiveness of various downstream neural network approaches using transformer-based models for COVID-19 detection. While previous research has shown the effectiveness of transformer-based models in detecting COVID-19-related news, limited research has explored how to combine these models with different downstream neural structures for COVID-19 detection. Thus, this study extends the existing research on transformer-based models by examining their effectiveness in the context of COVID-19 detection with different downstream neural structures. The study's findings are consistent with previous research on the effectiveness of transformer-based models, such as the ensemble transformer model, for COVID-19 detection. The CT-BERT+BiGRU model outperformed all other models, demonstrating the effectiveness of combining advanced ML models with PLMs in capturing context and generating informative representations for downstream tasks. However, the study's findings contrast with some previous research that suggested larger models with more parameters perform better. The study found that the base versions of the models had better accuracy than their larger counterparts, suggesting that adding more model parameters may lead to performance degradation. In addition, the study found that recurrent neural-based models outperformed CNN-based models, which contradicts some previous research that suggested CNN-based models are more effective in natural language processing tasks.

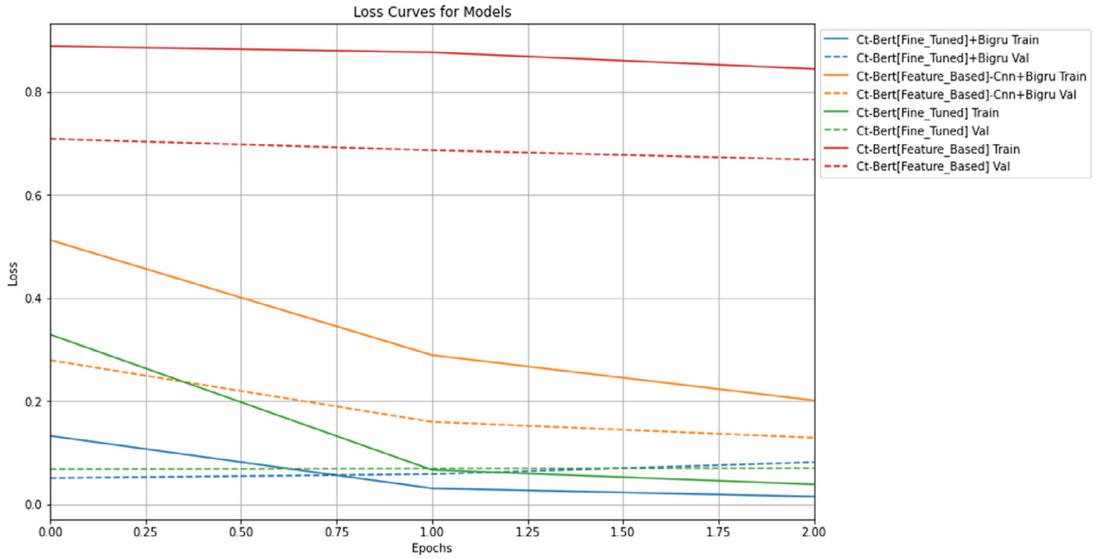


Fig. 5. Comparison (%) performance using *Loss* curves for models.

Table 7
A sample of misclassified classes obtained by CT-BERT+BiGRU model.

Tweet	Actual	Pred
A common question: are coronavirus cases going up because we are testing so many more people? A: Certainly not in Florida where testing slowed down 3% while new cases grew 88% over the last week.	Real	Fake
FDA official says if a COVID vaccine is approved before it is ready – he's outta there.	Real	Fake
Oxford coronavirus vaccine is safe and induces strong immune response early trial results suggest.	Fake	Real

We have the following observations.

- **Our hypothesis is valid.** This is observed from the outstanding performance of CT-BERT+BiGRU compared to the other baselines.
- **Adding different downstream neural structures on top of transformer-based models is effective.** This is observed through the superior performance of e.g., CT-BERT+BiGRU, compared to the original CT-BERT model.
- **Fine-tuning approach is effective.** This is observed from the promising results generated by fine-tuning the proposed models. This shows how effective the fine-tuning approach compared to the off-the-shelf approaches.

Although we have very interesting results in terms of recall, the precision of the model shows the portion of false detection we have. To better understand this phenomenon, we analyse the errors of the best-performing model. We investigate the confusion matrix resulting from the CT-BERT+BiGRU model shown in Fig. 6. It is evident that the model can separate fake from real content properly. Only ten samples belonging to the real class are misclassified as fake, and 23 of the fake samples are misclassified as real. Thus, almost 0.47% of real samples are misclassified as fake, while 1.07% of fake samples are classified as real. We provided some examples of misclassified samples in Table 7. We observed that the model is confused when classifying claims involving words like "vaccine".

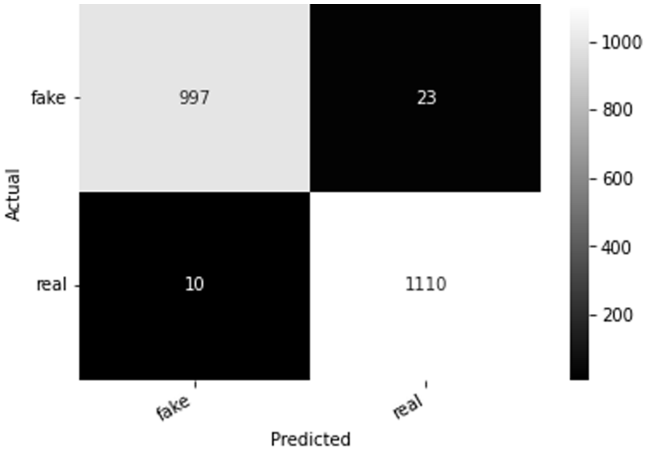


Fig. 6. Confusion Matrix of CT-BERT + BiGRU.

4.1. Limitations

The results presented in this study provide valuable insights into the effectiveness of various downstream neural network approaches for COVID-19 detection using transformer-based models. Nonetheless, future work could address some potential limitations. For example, the study did not explore the impact of different hyperparameters and optimisation techniques on model performance. Moreover, the study utilised a relatively small COVID-19 dataset, potentially limiting the generalisability of the findings. Further research could investigate the impact of larger datasets and the transferability of the proposed models to related tasks. In addition, examining the interpretability of the models could provide valuable insights into their underlying mechanisms and improve transparency and trustworthiness. Finally, the study found that CT-BERT+BiGRU outperforms other models, which is an interesting finding. However, it remains unclear whether this result is solely due to the pre-training task or if other factors are at play. Therefore, it is necessary to conduct further analysis to identify the specific factors that contribute to the proposed model's superior performance. Future research could investigate the impact of different pre-training tasks and datasets to determine the robustness of the proposed models.

5. Conclusions and future research

Our study explores the effectiveness of transformer-based models for COVID-19 fake news detection, presenting novel and effective approaches. Our findings indicate that transformer-based models outperform both traditional and advanced ML baselines for detecting COVID-19 fake news. Fine-tuned CT-BERT with BiGRU achieved state-of-the-art performance with an F1 score of 98.5%, highlighting the importance of fine-tuning and the potential of incorporating complex downstream neural structures. However, to fully understand the limitations of transfer learning and the behaviour of the detectors on different datasets with varying domains, further research is needed. This includes investigating the impact of hyperparameters and optimisation techniques, evaluating the robustness of the models against biased data, and improving interpretability. Overall, our study demonstrates the potential of PLMs for COVID-19-related fake news detection and has important implications for the development of more accurate fake news detection models.

Declaration of competing interest

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

Data availability

I have shared the link to the data.

References

[1] W.H. Organization, Statement on the second meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-nCoV), 2022, https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov), (Accessed June 20, 2022).

[2] K.-C. Yang, C. Torres-Lugo, F. Menczer, Prevalence of low-credibility information on twitter during the covid-19 outbreak, 2020, arXiv preprint arXiv:2004.14484.

[3] T.L. Huynh, et al., The COVID-19 risk perception: A survey on socioeconomics and media attention, Econ. Bull. 40 (1) (2020) 758–764.

[4] M.S. Islam, A.-H.M. Kamal, A. Kabir, D.L. Southern, S.H. Khan, S.M.M. Hasan, T. Sarkar, S. Sharmin, S. Das, T. Roy, M.G.D. Harun, A.A. Chughtai, N. Homaira, H. Seale, COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence, PLOS ONE 16 (5) (2021) 1–17, http://dx.doi.org/10.1371/journal.pone.0251605.

[5] C.M. Greene, G. Murphy, Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation, J. Exp. Psychol.: Appl. (2021).

[6] S. Busari, B. Adebayo, Nigeria records chloroquine poisoning after trump endorses it for coronavirus treatment, 2020.

[7] Poynter, Fighting the infodemic: The coronavirusfacts alliance, 2020, https://www.poynter.org/coronavirusfactsalliance/.

[8] J. Alghamdi, Y. Lin, S. Luo, Modeling fake news detection using BERT-CNN-BiLSTM architecture, in: 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), 2022, pp. 354–357, http://dx.doi.org/10.1109/MIPR54900.2022.00069.

[9] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 746–751.

[10] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[11] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N.A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8342–8360, http://dx.doi.org/10.18653/v1/2020.acl-main.740, Online, https://aclanthology.org/2020.acl-main.740.

[12] M. Müller, M. Salathé, P.E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020, arXiv preprint arXiv:2005.07503.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, 2019, ArXiv abs/1907.11692.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, ArXiv abs/1810.04805.

[15] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: International Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, Springer, 2021, pp. 21–29.

[16] F.A. Ozbay, B. Alatas, Adaptive salp swarm optimization algorithms with inertia weights for novel fake news detection model in online social media, Multimedia Tools Appl. 80 (26–27) (2021) 34333–34357, http://dx.doi.org/10.1007/s11042-021-11006-8.

[17] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. Pykl, A. Das, A. Ekbal, M.S. Akhtar, T. Chakraborty, Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts, in: International Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, Springer, 2021, pp. 42–53.

[18] E. Shushkevich, J. Cardiff, Tudublin team at constraint@aaai2021 - COVID19 fake news detection, 2021, ArXiv abs/2101.05701.

[19] S. Bandyopadhyay, S. Dutta, The analysis of fake news in social medias for four months during lockdown in COVID-19-a study: biostatistical analysis of COVID-19, Xeno J. Biomed. Sci. 1 (1) (2020) 1–6.

[20] T. Felber, Constraint 2021: Machine learning models for COVID-19 fake news detection shared task, 2021, arXiv preprint arXiv:2101.03717.

[21] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining, in: Ldv Forum, Vol. 20, Citeseer, 2005, pp. 19–62.

[22] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018.

[23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018.

[24] A. Aggarwal, A. Chauhan, D. Kumar, M. Mittal, S. Verma, Classification of fake news by fine-tuning deep bidirectional transformers based language model, EAI Endorsed Trans. Scalable Inf. Syst. 7 (27) (2020).

[25] H. Jwa, D. Oh, K. Park, J.M. Kang, H. Lim, exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert), Appl. Sci. 9 (19) (2019) 4062.

[26] A. Baruah, K.A. Das, F.A. Barbhuiya, K. Dey, Automatic detection of fake news spreaders using BERT, in: CLEF (Working Notes), 2020.

[27] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019, ArXiv abs/1910.01108.

[28] A. Gautam, V. V, S. Masud, Fake news detection system using xlnet model with topic distributions: Constraint@aaai2021 shared task, 2021, http://dx.doi.org/10.48550/ARXIV.2101.11425, https://arxiv.org/abs/2101.11425.

[29] B. Koloski, T.S. Perdih, M. Robnik-Šikonja, S. Pollak, B. Škrlj, Knowledge graph informed fake news classification via heterogeneous representation ensembles, Neurocomputing (2022).

[30] S.M.S.-U.-R. Shifath, M.F. Khan, M.S. Islam, A transformer based approach for fighting COVID-19 fake news, 2021, http://dx.doi.org/10.48550/ARXIV.2101.12027, https://arxiv.org/abs/2101.12027.

[31] Z. Gao, A. Feng, X. Song, X. Wu, Target-dependent sentiment classification with BERT, Ieee Access 7 (2019) 154290–154299.

[32] A. Nikolov, V. Radivchev, Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 691–695.

[33] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[34] P. Usherwood, S. Smit, Low-shot classification: A comparison of classical and deep transfer machine learning approaches, 2019, ArXiv abs/1907.07543.

[35] L. Horne, M. Matti, P. Pourjafar, Z. Wang, GRUBERT: A GRU-based method to fuse BERT hidden layers for Twitter sentiment analysis, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Suzhou, China, 2020, pp. 130–138, https://aclanthology.org/2020.aacl-srw.19.