Evaluating Pre-trained BERT-based Language Models for Detecting Misinformation

Rini Anggrainingsih (rini.anggrainingsih@staff.uns.ac.id)
Sebelas Maret University
Ghulam Mubashar Hassan
The University of Western Australia
Amitava Datta
The University of Western Australia

Research Article

Keywords: BERT, Fake News Detection, Finetuned Language Model, Rumor Detection

Posted Date: May 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1608574/v1

Evaluating Pre-trained BERT-based Language Models for Detecting Misinformation

Rini Anggrainingsih[1,2*], Ghulam Mubashar Hassan[2†] and Amitava Datta[2†]

[1*]Informatics Department of Mathematics and Natural Sciences Faculty, Sebelas Maret University, Ir Sutami 36A, Surakarta, 57126, Central Java, Indonesia.
[2]Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway, Perth, 6009, WA, Australia.

*Corresponding author(s). E-mail(s): rini.anggrainingsih@staff.uns.ac.id, rini.anggrainingsih@research.uwa.edu.au;
Contributing authors: ghulam.hassan@uwa.edu.au; amitava.datta@uwa.edu.au;
†These authors contributed equally to this work.

Abstract

It is challenging to control the quality of online information due to the lack of supervision. Manual checking is almost impossible given the vast number of posts made on online media and how quickly they spread. Therefore, there is a need for automated rumour detection techniques to limit the adverse effects of spreading misinformation. Previous studies mainly focused on finding and extracting the significant features of text data. However, extracting features is time-consuming and not a highly effective process. This study proposes the BERT-based pre-trained language models to encode text data into vectors and utilise neural network models to classify these vectors to detect misinformation. Furthermore, different language models (LM) ' performance with different trainable parameters was compared. The proposed technique is tested on different short and long text datasets. The result

of the proposed technique has been compared with the state-of-the-art methods on the same datasets. The results show that the proposed technique performs better than the state- of-the-art techniques. We also tested the proposed technique by combining the datasets. The results demonstrated that the large training and testing size of data considerably improves the technique's performance. Therefore, it is suggested that the dataset, splitting data, and classification techniques must be considered carefully to analyse performance of the solutions.

Keywords: BERT, Fake News Detection, Finetuned Language Model, Rumor Detection

1 Introduction

The rapid growth of internet technology and online social media has revolutionised communication. People can freely and easily access the latest news, share, and express their opinions on social media. However, these conveniences may promote the creation and the spread of false information, such as rumours and fake news that cause social problems, including financial loss, public panic, personal or community reputation defacement, and human life threats. Recently, spreading rumours and fake news on online media platforms has become a serious issue. It is nearly impossible for ordinary people to distinguish rumour and non-rumour information from most online platforms. Therefore, there is a need for automated detection technology to identify rumours or fake news on online social media.

Castillo et al. [1], who were known as pioneers in detecting credible information research area, used feature engineering and machine learning techniques to detect rumours on Twitter and provided a baseline result for other studies. Other studies then attempted to improve the accuracy of detecting misinformation by discovering the essential features from the context and content of text data. Furthermore, some other studies compared the performance of various machine-learning algorithms for detecting rumours, such as Random Forest, Support Vector Machine, K-Nearest Neighbour, Naïve Bayes Classifier, and Logistic Regression [2–5]. Similarly, feature engineering and machine learning techniques are also widely used to detect fake news on online media [6–9].

Following the success of the neural network based approaches in various classification problems, most recent studies leveraged different neural network based techniques for detecting rumours and fake news on online media, including Recurrent Neural Network (RNN) [10–12], Convolutional Neural Network (CNN) [13–17], and Long Short Term Memory (LSTM) network [18–20], statistical features fusion network [21], and propagation path aggregation network [22]. Some studies used different methodologies to detect rumours on Twitter, such as reinforcement learning [23], entity recognition and sentence configuration [24], and graph-based neural networks [25–29].

The emergence of Bidirectional Encoder Representation from Transformer (BERT) as a pre-trained language model (LM) in 2018 by Devlin et al. [30] has significantly improved the area of Natural Language Processing (NLP). BERT can understand a word's contextual meaning by considering words from the left and the right sides. Furthermore, it can be finetuned to deal with other NLP-related problems. Pre-trained LMs have been applied on various types of NLP tasks, such as text summarisation [31–33], text generation [34–36], and text classification [37–41]. BERT has become the current research trend and marks a new era of NLP because of its high-performance results. Despite the benefits and ability of BERT, some issues are attributed to its complexity. Thereofore, some studies attempted to improve the LM's performance based on BERT's architecture and introduced other LM, such as RoBERTa [42] and DistilBERT [43] which were computationally less complex.

Motivated by the ability of pre-trained LM to classify text, this study aims to examine and compare the output performance of three different BERT-based language models when combined with neural network classifiers to detect misinformation. Different LMs used are RoBERTa, BERT and DistilBERT. Each LM is finetuned on labelled datasets and encoded into vectors. Afterward, these vectors were used to train the neural network-based techniques to detect misinformation. The model was validated on short and long text datasets (Tweets and fake news) which are: Pheme1 and Pheme2 datasets from [3, 19], Twitter15 and Twitter16 datasets [10] and COVID-19 Fake News dataset from Mendeley [44]. The contributions of this study are:

- Proposing a new model to improve rumour detection performance based on combining finetuned LM and neural network.
- Thoroughly evaluating the performance of three different variants of the proposed model in detecting rumours and fake news on four different datasets.
- Introducing a new benchmark testing procedure where all datasets are distributed in the same manner for training, validation and testing.
- Demonstrating that the proposed model outperforms the recent state-of-the-art techniques in rumour and fake news detection.

The rest of the paper is structured as follows: Sections 2 and 3 review the related works on rumour detection and the details of the proposed model on utilising pre-trained LM and neural networks to detect rumour and fake news, respectively. The experiments are discussed to investigate the model's performance in Section 4. Finally, Section 5 summarises the study.

2 Related Work

This section explains spreading rumours and fake news on online social media. Later, recent studies related to both rumour and fake news detection techniques are categorised and discussed systematically in different subsections.

2.1 Rumour and Fake News Detection

Though studies give different definitions, rumor and fake news are often used interchangeably. A rumour is a story that seems credible but complicated to verify since some part of the information could remain unverified [45]. In contrast, fake news is false information that spreads and appears credible to gain biased public opinion on particular issues [46]. The lack of supervision over social media posts makes rumours and fake news propagate quickly in the society.

Spreading rumours and fake news has become a serious concern due to the fast growth of internet technology and online media platforms. These issues often lead to anxiety or scepticism and attract readers to share them without verification [3, 47]. Most rumour and fake news on different topics spread through online media daily and influence people's opinions. Therefore, there is an urgent need to detect rumours or fake news automatically since manual fact-checking is a challenging and time-consuming process.

2.2 Traditional Machine Learning-based Approaches

Most of the earlier studies regarding information credibility focused on finding and extracting the significant features from the sources, such as the text contents and user details [1, 2, 4–9, 48, 49]. Standard machine learning techniques, such as Random Forest, Conditional Random Field, Support Vector Machine, Naïve Bayes, and K-Nearest Neighbour were used to detect the authenticity of the tweets.

Similarly, researchers handled fake news detection problems. Recently, Al-Ahmad et al. conducted a comprehensive study. The study applied and compared selected evolutionary classification models, including Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), and Salp Swarm Algorithm (SSA), to detect fake news related to COVID-19 issues [9]. This study reported that the best results were obtained using K-Nearest Neighbor and Genetic Algorithm techniques.

Overall, feature extraction is a fundamental step of a machine learning approach. However, manually extracting features is a time-consuming process that becomes more challenging as some features are missing due to a user's security settings. This condition affects the effectiveness of the feature-based approach.

2.3 Deep Learning-based Approaches

Deep learning with neural network approaches has shown promising results in various text classification problems. In the context of false information detection, the widely implemented neural network framework approaches on detecting rumour are Recurrent Neural Network (RNN) [10–12], Long-Short Term Memory (LSTM) [18–20] and Convolutional Neural Network (CNN) [13–17]. Recent studies that used neural network-based hybrid approaches have

achieved the new state-of-the-art results, which are considered in this study as benchmark results and compared to the performance of the proposed methods.

Wu and Rao worked on rumour detection by utilising interaction between the features of rumours. The study proposed models called "Adaptive Interaction Fusion Networks (AIFN)" and "Gated Adaptive Interaction Networks (GAIN)" to identify true and false rumours on Twitter [50]. These techniques were used to identify interaction features among the tweets and capture semantic conflict in posts and comments. The study evaluated their models on Pheme2 dataset [19]. Later a "Decision Tree-based CoAttention model (DTCA)" was proposed to extend this study by utilising the interaction of credible comments as evidence to detect the truth of the tweet [51]. The study reported 82.46% accuracy and 82.5% F1-score.

Wang et al. [28] aimed to detect rumour by considering a background hidden knowledge in the post's text content and proposed a "Knowledge-driven Multimodal Graph Convolutional Network". This approach combined textual, conceptual and visual data to represent the semantic information into a unified framework for fake news identification. It used Pheme1 dataset [3] to validate the model. The study reported above 87% for all evaluation parameters including, accuracy, precision, recall and F1-score.

Lu and Li [27] aimed to predict whether the source of tweets was fake by using user profiles and social interactions. A Graph-Aware CoAttention Network (GCAN) was proposed based on neural network models to depict the interaction among users and capture the relationship between the source tweet, its propagation and user interaction to generate a prediction. The model was evaluated using Twitter15 and Twitter16 datasets [10]. Accuracy of 87% and 90% on Twitter15 and Twitter16 datasets were reported, respectively.

Wang and Guo encoded tweets' sentiment information and word representation with a two-layer Cascaded Gated Recurrent Unit (CGRU) to detect rumours [52]. The model was validated using Twitter16 datasets [10]. The model achieved 88.5% accuracy that outperformed earlier studies. Another study [15] classified tweets as rumour or non-rumour using a Dual Convolutional Neural Network (DCNN) technique using the inherent features of the information set. Their study used Twitter15 and Twitter16 datasets [10] to evaluate the model and reported F1 scores of 86% and 87%, respectively.

Zhanh et al. proposed a lightweight propagation path aggregating (PPA) neural network for rumor embedding and classification [22]. They first modeled the propagation structure of each rumor as an independent set of propagation paths in which each path represents the source post in a different conversation context. Then aggregated all paths to obtain the representation of the whole propagation structure. Twitter 15, Twitter 16 and Pheme1 datasets were used to evaluate their model and achieved accuracies of 87.3%, 88.7% and 80.3% for Twitter15, Twitter16 and Pheme1, respectively.

Another model proposed by Ma et al.[24] by integrating entity recognition, sentence reconfiguration, and ordinary differential equation network under a unified framework called ESODE. They used an entity recognition method to

enhance the semantic understanding of rumor texts, and designed a sentence reconfiguration to improve the frequency of important words, then established a complete feature map by collecting statistical features from three aspects, including linguistic features on the content of rumors, characteristics of users involved in rumor propagating, and propagation network structures. They reported results on Twitter15-16 dataset by achieving 92.4% precision, 92.6% of recall, and 92.5% for F1-score.

2.4 BERT-based Language Model

The emergence of Language Models (LM) and transformers significantly improved the solutions related to Natural Language Processing (NLP). BERT is a pre-trained language model trained on vast unlabelled data from the BooksCorpus that includes 800 million words and English Wikipedia with 2,500 million words without any genuine training objective. Therefore, it can be finetuned for many NLP tasks. BERT can understand a word's contextual meaning by considering words from both the left and the right sides. Furthermore, it can represent words and sentences that are converted into numeric vectors [30]. There are two standard architectures of BERT: $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$. In general, $\text{BERT}_{\text{BASE}}$ has 12 encoder layers with 110 million parameters and 768 hidden layers, and $\text{BERT}_{\text{LARGE}}$ has 16 encoder layers with 340 million parameters and 1,024 hidden layers.

Despite the benefit and ability of the pre-trained language model, there are issues regarding training, memory consumption, and computational power due to a large amount of training data. Therefore, some researchers introduced different LMs to address these issues, such as RoBERTa [42] to address a training model issue and DistilBERT [43] to deal with the memory and computational inefficiency problems.

Liu et al. examined the effect of hyperparameter tuning and training size on BERT architecture [42]. The results showed that BERT was significantly under-trained. Therefore, an improved BERT training model called Robustly Optimised BERT Approach (RoBERTa) was proposed. The training procedures were modified, including; 1) training the model with a larger dataset; 2) removing the next sentence prediction objective; 3) training in a more extended sequence; and 4) changing the masking pattern dynamically to the training data. RoBERTa was trained by following $\text{BERT}_{\text{LARGE}}$ architecture. However, unlike BERT, which initially trained using 16GB of the dataset, a more extensive training dataset containing 160GB of text was used to train RoBERTa. In addition, RoBERTa was trained for more iterations of 300,000, which was later extended to 500,000. RoBERTa consistently outperformed BERT in all individual tasks on the standard GLUE benchmark [53].

Although BERT and RoBERTa, are leading in the current NLP research, the large models raise some issues regarding the computational cost and requirements. Sanh et al. proposed a lighter and faster LM based on BERT architecture using knowledge distillation to compress the model to deal with computational issues [43]. Furthermore, they trained the compressed model

to reproduce the behaviour of the large model. The compressed model is called DistilBERT. It has 66 million parameters, 40% fewer than BERT and trains 60% faster than BERT. The results showed that DistilBERT models could achieve good results on several NLP tasks and are computationally light enough to run on mobile devices.

Some researchers in the information credibility area who applied BERT-based LMs in classifying misinformation have reported that they obtained an excellent performance by using LMs instead of using standard machine learning methods [54–56]. Furthermore, other researchers compared LMs' performance to evaluate the cross-source failure problem in the current misinformation detection methods. They focused on finding generalisable representation so that the classification model would be more applicable in real-world data [57]. They examined the cross-source generalisability by choosing one dataset as a training set and treating other datasets as testing sets.

Pelrine et al. evaluated the performance of a few pre-trained Language Models on detecting rumours [58]. However, they took a different direction to facilitate a solid standard benchmark by using different distribution of data on each dataset just same with the previous studies to make a fair benchmark comparison. The LMs were finetuned, and a single layer perceptron was used as a classifier in their study. The results showed that the performance of the proposed method was better than the state-of-the-art models. Therefore, Pelrine's et al. study is included as a current state-of-the-art model and compared with the proposed methods.

3 Material and Method

This section describes the proposed models and datasets used to validate the proposed models and explains the experimental steps used in this study. This study focuses on comparing and evaluating the performance of three different size BERT-based LMs, including RoBERTa, $\text{BERT}_{\text{BASE}}$, and DistilBERT on distinguishing rumour and non-rumour tweets, classifying true and false-rumours, and detecting true and fake-news by using neural network-based classifier. Various datasets were then used to validate the performance of the proposed methods. The proposed method and datasets are explained below.

3.1 The Proposed Model

This study utilises the finetuned model of LMs as an encoder to represent text into vectors and applying neural network models as a classifier. The architecture of the proposed method is presented in Figure 1.

The finetuned LM was leveraged to encode the sentences into vectors. After the data was encoded into vectors by finetuned LMs, two different neural network techniques were used for classification, including Multilayer Perceptron (MLP) and Residual Network on Convolutional Neural Network (Resnet-CNN). MLP is a neural network that comprises more than one perceptron.

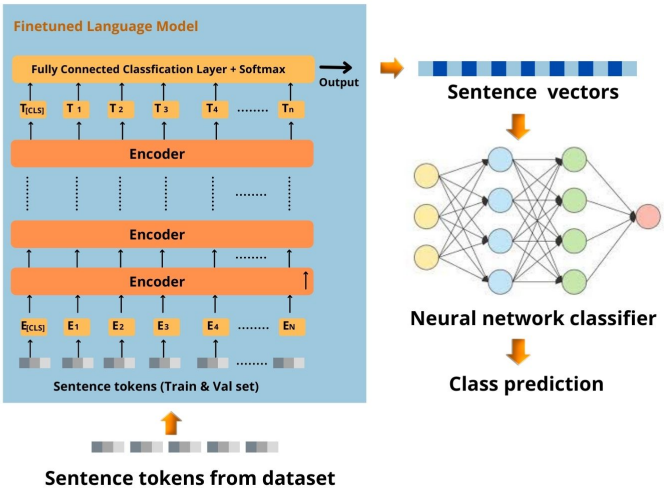

Fig. 1 The proposed architecture for detecting credibility of information

Generally, it consists of an input layer to receive the signal, an output for predicting the input, and an arbitrary number of hidden layers between the input and output [59]. This study uses four layers of MLP (4L-MLP) and performed regularisation and dropout on the 4L-MLP model (4L-MLP+Reg+Drop). While ResNet-CNN is an improved model for addressing the vanishing gradient problem that appears on a Convolutional Neural Network (CNN) with deeper layers. A CNN is similar to MLP, though it comprises several layers in its hidden layer, including convolutional, pooling, fully connected, and regularisation layers. These additional layers play an essential role in the success of most deep neural networks. However, a study [60] empirically showed a maximum threshold for the depth of the additional layers of the CNN model. A new neural layer called the residual network was introduced to deal with the problem of deeper networks. ResNet comprises residual blocks, which skip connections to some layers and join the outputs from prior layers to the output of stacked layers. This skipped connections could solve the vanishing gradient problem in CNN and improve neural networks' performance with more layers [60]. This study uses 10-layers and 18-layers ResNet-CNN for classification purposes.

3.2 Datasets

Both long-text (fake news) and short-text (tweets) datasets were used to validate the proposed method's performance. Regarding the splitting data technique, this study suggests to take the same distribution of data for all datasets to make a standard testing procedure. It is different from many existing studies, that took a different distribution based on the state-of-the-art results of each dataset. The datasets are explained in detail in the following subsections.

3.2.1 COVID-19 Fake News Dataset

COVID-19 Fake News Dataset from Mendeley [44] was selected to represent a long-text dataset. This dataset contains news about COVID-19 issues from December 2019 to July 2020. Each item is labelled as true-news (2,061) and as false-news (1,058). It is called as Covid T/F dataset.

3.2.2 Pheme Datasets

There are two types of Pheme datasets: Pheme1 and Pheme2. Pheme1 contains 5,791 tweets about five breaking news topics where 1,969 tweets are labelled as rumours and 3,822 are labelled as non-rumours by journalists [3]. Pheme2 is the extended version of Pheme1, with four additional news topics, meaning that Pheme2 contained 6,425 tweets about nine news events in total. The journalists verified whether these tweets are rumour (2,402 tweets) or non-rumour(4,023 tweets). Then, rumour tweets are further classified and labelled as true-rumour (1,067 tweets), false-rumour (638 tweets) or unverified-rumour (697 tweets) [19]. However, this study does not consider the data with unverified-rumour label as it is beyond the scope of this study.

Pheme datasets are used in different manner in different state-of-the-art techniques. Therefore we evaluated the proposed models in the same manner to make a fair comparison with existing studies. These detailed sub-datasets are explained below:

- Pheme1 R/NR: It contains 5,791 tweets where 1,969 are labelled as rumour tweets and 3,822 are labelled as non-rumour tweets from Pheme1 dataset.
- Pheme2 T/F: It contains only 1,705 tweets that are marked as true or false rumours and taken from Pheme2 dataset. It consists of 1,067 and 638 tweets labelled as true-rumours and false-rumours, respectively.
- Pheme2 R/NR: It consists of 6,425 tweets from Pheme2 dataset where 2,402 tweets are categorised as rumour tweets while 4,023 are categorised as non-rumour tweets.

3.2.3 Twitter15 and Twitter16

Twitter 15 and 16 are publicly available datasets containing 1,490 and 818 tweets. The journalists labelled each tweet as rumour or non-rumour. Afterwards, each rumour tweet is classified and labelled as a true-rumour, false-rumour or unverified rumour [10]. Similar to the Pheme2 dataset, this study did not consider the data with unverified-rumour labels. This study uses four versions of this datasets, as follows:

- Twitter15 R/NR: It contains 1,490 tweets from the Twitter15 dataset where 1,118 are labelled as a rumour while 372 are labelled as non-rumour.
- Twitter16 R/NR: It consists of 818 tweets from the Twitter16 dataset marked as a rumour (613 tweets) or non-rumour (205 tweets).
- Twitter15 T/F: It contains tweets labelled as true-rumour (375 tweets) or false-rumour (370 tweets) from the Twitter15 dataset.

- Twitter16 T/F: It comprises 410 tweets from Twitter16 dataset, which are marked as true-rumour (205 tweets) or false-rumour (205 tweets).

3.2.4 Combined Dataset

In order to generalise the performance of the proposed method, we combined the dataset as suggested by [58]. In this study, we approached the problem differently than [57] where different datasets are used for training and testing. This study generalised combined short-text datasets with the same label from Pheme2, Twitter15, and Twitter16 datasets to obtain a generalised performance result in classifying rumour and non-rumour tweets and, distinguishing true and false rumour tweets.

A single combined dataset with rumour and non-rumour label named Combined R/NR was created from Pheme2 R/NR, Twitter15 R/NR, and Twitter16 R/NR to validate the model's performance in classifying rumour and non-rumour tweets. Hence, Combined R/NR dataset comprises 4,600 non-rumour tweets and 4,133 rumour tweets. Similarly, Pheme2 T/F, Twitter15 T/F, and Twitter16 T/F datasets were incorporated into a single dataset named Combined T/F dataset. The dataset consists of 1,213 false-rumour tweets and 1,646 true-rumour tweets. Table 1 describes the more detailed distribution of the datasets used in this study.

Table 1 Distribution of labels in the datasets used in this study

Datasets and labels	COVID-19 Fake News [44]	Pheme1 [3]	Pheme2 [19]	Twitter 15 [10]	Twitter 16 [10]	Combined R/NR	Combined T/F
False-news	1,058	-	-	-	-	-	-
True-news	2,061	-	-	-	-	-	-
Non-rumours	-	3,822	4,023	372	205	4,600	-
Rumour	-	1,969	2,402	1,118	613	4,133	-
False-rumours			638	370	205	-	1,213
True-rumours			1,067	374	205	-	1,646
Unverified (not used)			*697*	*374*	*203*	-	-

3.3 Experimental Procedure

The experiments used a RTX 2080 GPU to train each model. The experimental setup for identifying rumours/non-rumours tweets, and true/fake-news using LMs and neural network models is shown in Figure 2.

As a first step, dataset was split into a training, validation, and testing sets. From each dataset, 10% of the data is reserved first for testing. The remaining data was split into 75% and 25% for training and validation sets, respectively. For the combined datasets, 10% of the data was taken from each

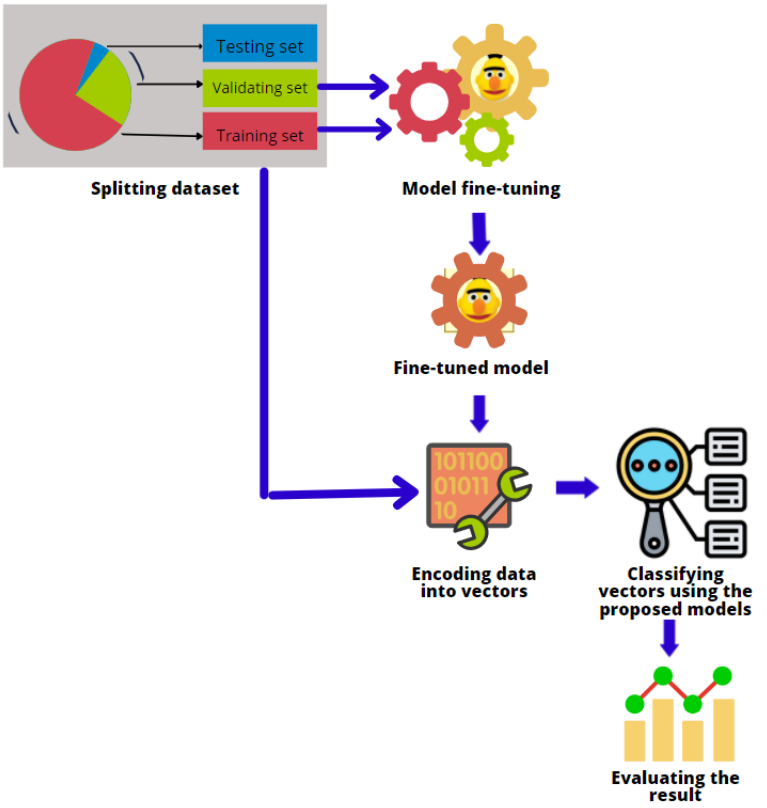

Fig. 2 The experiment setup to identify rumours/non-rumour tweets and true/fake news using the proposed model.

dataset for the testing-set before blending them into a single Combined T/F dataset or Combined R/NR dataset for training and validation. The rest of combined dataset was split into 3:1 proportions for the training and validation sets, respectively.

The next step was the finetuning of the LM using Huggingface library [61] and the labelled data as an input. Adam optimiser was used while the learning rate was set at 5e-5 with a batch size of eight and experiments were run for ten epochs. The mean pooling layer was set as a pooler to encode the datasets into vectors. Primarily, these vectors were used to train the classifier model.

In the classification step, the learning rate was set at 2e-4, batch size at 512, and the number of maximum epochs selected was 1,000. The Models were trained using Adam optimiser, and cross-entropy as the loss functions. At the end, the models were evaluated by calculating their accuracy, precision, recall, and F1-score using equations(1) to (4):

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

$$Precision(P) = \frac{TP}{TP + FP} \tag{2}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{3}$$

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

where TP represents True-positive cases, FN represents False-negative cases, FP represents False-positive cases, and TN represents True-negative cases.

4 Results and discussion

This section presents the experimental results of the proposed models and compares them with the results of state-of-the-art models for each dataset. The works which are considered as the state-of-the-art models on each dataset and their performance are presented in Table 2. The comparative performance of all compared models are presented in Table 3 to Table 10. Each table's first and second rows provide details of the state-of-the-art models and their splitting data strategy that achieved the state-of-the-art performance. The best result from each column is in bold, and the second-best and the third-best results are represented in underlined and italic font, respectively.

4.1 Comparison of Results with State-of-the-art Methods

Table 3, Table 4, and Table 5 present the experimental results of the proposed methods on classifying rumour or non-rumour on different datasets. Table 3 presents the experimental results and their comparison in distinguishing rumour and non-rumour tweets on the Pheme1 R/NR dataset. It can be observed that all the proposed models using finetuned RoBERTa LM outperformed the state-of-the-art technique of [28]. The best scores achieved are 88.9% accuracy, 87.9% precision, 87.7% recall and 87.8% F1-score. Although the proposed models attained slightly lower F1-score than [58] while other metrics were not available for [58].

Table 4 illustrates the experimental results of the proposed models on Pheme2 R/NR dataset that distinguishes rumour and non-rumour tweets. The

Table 2 State-of-the-art results for rumour and fake news detection for each dataset.

Datasets	Sources	Splitting data strategy	Best performance			
			A (%)	P (%)	R (%)	F1 (%)
Pheme1 R/NR	[28]	70:10:20 for train:val:test	87.56	87.62	87.65	87.64
	[58]	70:10:20 for train:val:test	-	-	-	89.4±0.3
Pheme2 R/NR	[7]	Not provided	83.36	83.59	99.64	90.91
Pheme2 T/F	[51]	70:10:20 for train:val:test	82.46	79.08	86.24	82.5
	[58]	70:10:20 for train:val:test	-	-	-	93.2±0.9
Twitter15 T/F	[27]	70:30 for train:test	87.67	82.57	82.95	82.50
	[58]	70:10:20 for train:val:test	-	-	-	94.4±0.8
Twitter16 T/F	[27]	70:30 for train:test	90.84	75.94	76.32	75.93
	[58]	70:10:20 for train:val:test	-	-	-	95.7±2.8
Twitter15 R/NR	[62]	Not provided	86.3	-	-	-
	[15]	60:40 for train:test	86	-	-	-
Twitter16 R/NR	[52]	75:25 for train:test	88.5	-	-	-
	[15]	60:40 for train:test	87	-	-	-
Covid T/F	[9]	Not provided	75.43	66.22	56.33	60.9

Abbreviations: A:Accuracy, P: Precision, R: Recall, F1: F1 score

observations are similar to Pheme1 R/NR results, where RoBERTa performed generally better than other LMs. This can be contributted to the fact that RoBERTa has higher number of parameters than other LMs. It can also be observed that 4L-MLP with regularisation and dropout, outperformed both ResNet-CNN classifiers. However, the difference in classifier's performance between RoBERTa, BERT, and DistillBERT on the Pheme2 R/NR dataset is small, and it only varies from 1% to 3%. In summary, all the proposed models outperform the state-of-the-art model in accuracy and precision, while are similar on F1 score.

Table 5 compares the proposed models' performance on Twitter15 R/NR and Twitter16 R/NR datasets. The results show that all the proposed models perform exceptionally well on Twitter15 R/NR and Twitter16 R/NR datasets. The proposed models using finetuned RoBERTa and BERT LMs outperform the state-of-the-art models ([15, 28, 62]) on Twitter15 R/NR dataset. Interestingly on the Twitter16 R/NR dataset, RoBERTa LM which has highest

Table 3 Comparison of the proposed method with state-of-the-art techniques for Pheme1 R/NR dataset

Models		Best results on Pheme1 R/NR (5,791 tweets; 1,969 R, 3,822 NR)			
SOTA models and their split of dataset		A (%)	P (%)	R (%)	F1 (%)
[28]	70:10:20 for train:val:test	87.56	87.62	87.65	87.64
[58]	70:10:20 for train:val:test	-	-	-	89.4±0.3
Proposed Models [10%:testing, 75% and 25% of rest of the data for training and validation respectively]					
Classifiers	Language models	A (%)	P (%)	R (%)	F1 (%)
4MLP	BERT	87.304	86.416	85.448	85.929
	RoBERTa	88.215	*86.862*	87.643	87.251
	DistilBERT	87.130	86.258	85.201	85.726
4L-MLP + Reg+Drop	BERT	87.040	86.017	84.684	85.345
	RoBERTa	**88.908**	**87.900**	**87.728**	**87.814**
	DistilBERT	87.130	86.184	85.314	85.747
10L- ResNet-CNN	BERT	87.840	86.612	86.093	86.352
	RoBERTa	*88.000*	87.215	85.640	86.420
	DistilBERT	87.130	86.337	85.087	85.707
18-L ResNet-CNN	BERT	87.040	86.017	84.684	85.345
	RoBERTa	87.868	86.755	*86.588*	*86.671*
	DistilBERT	87.652	86.808	85.830	86.316

Abbreviations:
A:Accuracy, P: Precision, R: Recall, F1: F1 score
SOTA: State-of-the-art

number of trainable parameters obtained the worst accuracy, while DistilBERT LM which has smallest number of trainable parameters achieved the best accuracy. It is expected that size and imbalance of the dataset may have affected the performance of LMs.

For classifying true and false rumours, Tables 6 and 7 presents the performance of the proposed methods on classifying true and false rumours using Pheme2 T/F, Twitter15 T/F, and Twitter16 T/F datasets. It can be observed from the tables that these datasets are relatively small in size but are well balanced. Table 6 presents the experimental results of the proposed models on Pheme2 T/F dataset for identifying true and false rumours in tweets. All the proposed models outperform the state-of-the-art model [51] in all metrics while similar to [58] in terms of F1 score. This shows that all the proposed models consistently achieve a high performance on Pheme2 T/F dataset. Interestingly, by using 4-layers MLP and ten-layers ResNet-CNN classifier models, BERT and DistilBERT LMs obtained better result as compared to RoBERTa LM which has more trainable parameters. Table 7 compares the proposed models' performance on classifying true and false-rumour on Twitter15 T/F and Twitter16 T/F datasets. It can clearly observed that all the proposed models

Table 4 Comparison of the proposed method with state-of-the-art techniques for Pheme2 R/NR dataset

Classifier models		Best results on Pheme2 R/NR (6425 tweets; 2402 R, 4023 NR)			
SOTA models and their split of dataset		A (%)	P (%)	R (%)	F1 (%)
[63]	Not provided	83.36	83.59	99.64	90.91
Proposed Models [10%:testing, 75% and 25% of rest of the data for training and validation respectively]					
4MLP	BERT	89.404	88.812	88.532	88.672
	RoBERTa	*90.066*	*89.280*	89.676	*89.478*
	DistilBERT	86.342	85.553	85.983	85.767
4L-MLP + Reg+Drop	BERT	88.907	88.156	88.222	88.189
	RoBERTa	**90.232**	89.518	**89.721**	**89.61**
	DistilBERT	86.424	85.819	85.006	85.411
10L-ResNet-CNN	BERT	89.073	88.403	88.267	88.335
	RoBERTa	89.735	88.895	89.411	89.152
	DistilBERT	87.252	86.478	86.282	86.380
18-L ResNet-CNN	BERT	89.735	89.227	88.797	89.011
	RoBERTa	90.232	**89.565**	*89.634*	89.599
	DistilBERT	87.086	86.279	86.150	86.214

Abbreviations:
A:Accuracy, P: Precision, R: Recall, F1: F1 score
SOTA: State-of-the-art

achieve high performance and outperform the state-of-the-art models ([27, 58]) on both datasets. For Twitter15 T/F dataset, BERT LM with 4-layers MLP classifier performs the best while for Twitter16 T/F dataset, RoBERTa LM with the same classifier obtained the best scores. Therefore, it can be concluded that MLP as a classifier performs better than all other considered classifiers.

The experimental results of the proposed models for classifying true and fake news on Covid T/F as a long-text dataset are presented in Table 8. The results indicate that all the proposed models perform better on Covid T/F dataset as compared to the State-of-the-art methods [9]. The models' performance considerably improves the accuracy by around 5% - 7% and F1-Score by 16%-19% as compared to the state-of-the-art technique. RoBERTa as the largest LM combined with a simple 4-layers MLP classifier performs better than other LMs and classifiers. However, when incorporated with a more complex classifier, such as 18L-Resnet-CNN, RoBERTa's performance slightly decreased. In general, there is no significant difference between the performance of three LMs.

4.2 Classification on Combined Datasets

Table 9 and Table 10 present the result on distinguishing rumour/non-rumour and true/false-rumour on combined datasets. As mentioned earlier, this study combined all the same labelled datasets from different sources into a new single

Table 5 Comparison of the proposed method with state-of-the-art techniques for Twitter15 R/NR and Twitter16 R/NR datasets

Classifier models		Best results on Twit15 R/NR (1490 tweets; 1118 R, 372 NR)				Best results on Twitter16 R/NR (818 tweets; 613 R, 205 NR)			
SOTA and split of dataset		A (%)	P (%)	R (%)	F1 (%)	A (%)	P (%)	R (%)	F1 (%)
[62]	Not provided	86.3	-	-	-	88.5	-	-	-
[15]	60:40 for train:test	86	-	-	-	87	-	-	-
Proposed Models [10%:testing, 75% and 25% of rest of the data for training and validation respectively]									
4MLP	BERT	87.586	82.711	82.711	82.711	*91.358*	90.211	**85.861**	**87.982**
	RoBERTa	92.466	89.769	88.944	89.355	85.294	83.103	73.077	77.768
	DistilBERT	85.616	81.155	76.287	78.646	**91.358**	**92.121**	84.180	87.972
4MLP +Reg +Dropout	BERT	88.966	87.266	80.551	83.774	90.123	91.205	81.680	86.18
	RoBERTa	*92.466*	90.505	*87.920*	*89.194*	86.420	82.767	79.221	80.955
	DistilBERT	85.616	80.266	78.335	79.289	91.358	92.121	*84.180*	*87.972*
10L- ResNet -CNN	BERT	87.586	85.882	77.610	81.537	90.123	91.205	81.680	86.18
	RoBERTa	**93.151**	**91.118**	**89.391**	**90.246**	86.420	83.939	77.541	80.613
	DistilBERT	86.986	82.803	79.228	80.976	91.358	*92.121*	84.180	87.972
18L -ResNet -CNN	BERT	86.538	82.181	81.624	81.902	90.123	91.205	81.680	86.18
	RoBERTa	92.466	*90.505*	87.920	89.194	87.654	85.240	80.041	82.559
	DistilBERT	88.356	83.539	84.217	83.877	91.358	92.121	84.180	87.972

Abbreviations:
A:Accuracy, P: Precision, R: Recall, F1: F1 score
SOTA: State-of-the-art

dataset to attain a larger and more balanced dataset. It is expected that large data will help to generalise and improve the training of the proposed models. The testing was done with testing set of each individual dataset which was separated before combining the data for training and validation, as well as with combined testing sets taken from all datasets. Therefore, Table 9 compares the models' performance on four different datasets, including Twitter16 R/NR, Twitter15 R/NR, Pheme1 R/NR, Pheme2 R/NR and their combination, and Table 10 compares the models' performance on four different datasets including Twitter16 T/F, Twitter15 T/F, Pheme2 T/F, and their combination.

It can be observed from the results presented in Table 9 and Table 10 that the proposed models perform well and obtain high performance as compared to the results obtained on individual datasets presented in Tables 3-8. This is due to the fact that more training and validation data is made available for LMs and classifiers. There was no significant difference in performance for light or large LMs with a simple or more complex classifier. The difference between all the models on combined dataset was around 1% to 2%. This shows that combining data can increase generalisation during the training process and help stabilise the proposed model's performance.

Table 6 Comparison of the proposed method with state-of-the-art techniques for Pheme2 T/F dataset

Classifier models		Best results on Pheme2 T/F (1705 tweets; 1067 T, 638 F)			
SOTA techniques and their split of dataset		A (%)	P (%)	R (%)	F1 (%)
[51]	70:10:20 for train:val:test	82.46	79.08	86.24	82.5
[58]	70:10:20 for train:val:test sets	-	-	-	93.2 $\pm$0.9
Proposed Models [10%:testing, 75% and 25% of rest of the data for training and validation respectively]					
4MLP	BERT	90.419	90.286	**89.179**	*89.729*
	RoBERTa	89.873	89.522	*89.015*	89.268
	DistilBERT	90.323	**92.113**	86.866	*89.413*
4L-MLP + Reg+Drop	BERT	89.820	89.782	88.385	89.078
	RoBERTa	89.241	89.251	87.891	88.566
	DistilBERT	89.820	90.544	87.759	89.13
10L-ResNet-CNN	BERT	**91.398**	91.840	89.116	**90.457**
	RoBERTa	88.623	90.162	85.859	87.958
	DistilBERT	89.241	88.741	88.499	88.62
18-L ResNet-CNN	BERT	*90.419*	*90.614*	88.866	89.731
	RoBERTa	89.241	88.741	88.499	88.62
	DistilBERT	87.425	87.140	85.836	86.483

Abbreviations:
A:Accuracy, P: Precision, R: Recall, F1: F1 score
SOTA: State-of-the-art

Table 7 Comparison of the proposed method with state-of-the-art techniques for Twitter15 T/F and Twitter16 T/F datasets

Classifier models		Best results on Twitter15 T/F (754 tweets; 374 T, 370 F)				Best results on Twitter16 T/F (410 tweets; 205 T, 205 F)			
State-of-the-art techniques and their split of dataset		A (%)	P (%)	R (%)	F1 (%)	A (%)	P (%)	R (%)	F1 (%)
[27]	70:30 for train:test	87.67	82.57	82.95	82.50	90.84	75.94	76.32	75.93
[58]	70:10:20 for train:val:test	-	-	-	94.4$\pm$0.8	-	-	-	95.7$\pm$2.8
Proposed Models [10%:testing, 75% and 25% of rest of the data for training and validation respectively]									
4MLP	BERT	**97.015**	**97.059**	**97.143**	**97.101**	92.308	93.182	92.500	92.840
	RoBERTa	97.015	*97.009*	*97.009*	*97.009*	**97.436**	**97.500**	**97.500**	**97.500**
	DistilBERT	95.522	95.499	95.580	95.539	88.889	88.889	88.259	88.573
4L-MLP + Reg+Drop	BERT	95.000	95.455	95.000	95.227	92.308	93.182	92.500	92.840
	RoBERTa	*97.015*	97.009	97.009	97.009	97.436	97.500	97.500	97.500
	DistilBERT	94.030	94.018	94.018	94.018	88.889	88.889	88.259	88.573
10L-ResNet-CNN	BERT	95.522	95.714	95.714	95.714	94.872	95.455	94.737	95.095
	RoBERTa	97.015	97.009	97.009	97.009	97.436	*97.500*	*97.500*	*97.500*
	DistilBERT	94.030	94.018	94.018	94.018	*97.436*	97.619	97.368	97.493
18-L ResNet-CNN	BERT	92.537	93.243	92.857	93.05	94.872	94.868	94.868	94.868
	RoBERTa	95.522	95.499	95.580	95.539	97.436	97.500	97.500	97.500
	DistilBERT	97.015	97.059	97.143	97.101	88.889	88.889	88.259	88.573

Abbreviations: A:Accuracy, P: Precision, R: Recall, F1: F1 score

Table 8 Comparison of the proposed method with state-of-the-art techniques for Covid T/F dataset

Classifier models		Best results on Covid T/F (3119 News; 2061 T, 1058 F)			
State-of-the-art techniques ([9]) Splitting data strategies are not provided		A (%)	P (%)	R (%)	F1 (%)
k-NN-BSSA		72.610	59.620	59.740	59.68
k-NN-BPSO		73.120	61.980	53.780	57.59
k-NN-BGA		75.430	66.220	56.330	60.88
k-NN		70.650	56.420	59.360	57.85
J48		72.290	59.600	56.900	0.58.22
RF		70.410	63.420	30.150	40.87
SVM		70.750	56.650	58.790	57.7
Proposed Models [10%:testing, 75% and 25% of rest of the data for training and validation respectively]					
4MLP	BERT	80.192	78.640	77.024	77.824
	RoBERTa	82.026	*79.961*	79.518	79.739
	DistilBERT	80.511	79.612	76.438	77.993
4L-MLP + Reg+Drop	BERT	81.046	79.223	77.345	78.273
	RoBERTa	**82.353**	**80.549**	79.286	**79.913**
	DistilBERT	80.511	*79.788*	76.229	77.968
10L-ResNet-CNN	BERT	81.046	79.223	77.345	78.273
	RoBERTa	81.046	78.730	79.258	78.993
	DistilBERT	80.511	79.982	76.021	77.951
18-L ResNet-CNN	BERT	*81.699*	79.474	**79.750**	*79.612*
	RoBERTa	81.046	78.722	*79.497*	79.108
	DistilBERT	80.192	79.679	75.567	77.569

Abbreviations: A:Accuracy, P: Precision, R: Recall, F1: F1 score

Table 9 Comparison of the proposed method with State-of-the-art techniques for Twitter16 R/NR, Twitter15 R/NR, Pheme1 R/NR, Pheme2 R/NR and Combined R/NR datasets

Classifier models		Twitter16 R/NR (613/205)		Twitter15 R/NR (1,118/372)		Pheme1 R/NR (1,969/3,822)		Pheme2 R/NR (2,402/4,023)		Combined R/NR (4,133/4,600)	
Approaches	LM	A%	F1 (%)	A (%)	(F1%)	A (%)	F1 (%)	A (%)	F1 (%)	A (%)	F1 (%)
4MLP	BERT	91.358	**87.982**	87.586	82.711	88.215	87.251	*90.066*	*89.478*	88.449	88.479
	RoBERTa	85.294	77.768	92.466	89.355	87.304	85.929	89.404	88.672	**90.099**	90.092
	DistilBERT	**91.358**	87.972	85.616	78.646	87.130	85.726	86.342	85.726	88.999	88.991
4L-MLP+ Reg+Drop	BERT	90.123	86.18	88.966	83.774	**88.908**	**87.814**	**90.232**	**89.613**	88.779	88.831
	RoBERTa	86.420	80.955	*92.466*	*89.194*	87.040	85.345	88.907	88.189	90.099	*90.091*
	DistilBERT	*91.358*	*87.972*	85.616	79.289	87.130	85.747	86.424	85.411	89.109	89.101
10L-ResNet-CNN	BERT	90.123	86.180	87.586	81.537	*88.000*	86.420	89.735	89.152	88.889	88.945
	RoBERTa	86.420	80.613	**93.151**	**90.246**	87.840	86.352	89.073	88.335	*90.099*	**90.107**
	DistilBERT	91.358	87.972	86.986	80.976	87.130	85.707	87.252	86.380	89.219	89.210
18-L ResNet-CNN	BERT	90.123	86.180	86.538	81.902	87.868	*86.671*	90.232	89.599	88.559	88.593
	RoBERTa	87.654	82.559	92.466	89.194	87.040	85.345	89.735	89.011	90.096	90.090
	DistilBERT	91.358	87.972	88.356	83.877	87.652	86.316	87.086	86.214	88.991	88.991

Abbreviations: A:Accuracy, F1: F1 score

5 Conclusion

This study demonstrates that incorporating a BERT-based finetuned language model as an encoder and a neural network as a classifier can improve the accuracy of misinformation detection. The proposed models consistently achieved

Table 10 Comparison of the proposed method with state-of-the-art techniques for Twitter16 T/F, Twitter15 T/F, Pheme T/F and Combined T/F datasets

Classifier models		Twitter16 T/F (410 tweets)		Twitter15 15 T/F (754 tweets)		Pheme2 T/F (1705 tweets)		Combined T/F (2869 tweets)	
Approaches	LMs	A%	F1 (%)	A (%)	F1 (%)	A (%)	F1 (%)	A (%)	F1 (%)
4MLP	BERT	92.308	92.840	**97.015**	**97.101**	90.419	*89.729*	**93.624**	**93.567**
	RoBERTa	**97.436**	**97.500**	97.015	97.009	89.873	89.268	92.617	92.459
	DistilBERT	88.889	88.573	95.522	95.539	*90.323*	89.413	92.883	92.814
4L-MLP+ Reg+Drop	BERT	92.308	92.840	95.000	95.227	89.820	89.078	93.624	93.567
	RoBERTa	97.436	97.500	*97.015*	*97.009*	89.241	88.566	92.282	92.124
	DistilBERT	88.889	88.573	94.030	94.018	89.820	89.130	92.953	92.778
10L-ResNet-CNN	BERT	94.872	95.095	95.522	95.714	**91.398**	**90.457**	92.617	92.421
	RoBERTa	97.436	*97.500*	97.015	97.009	88.623	87.958	92.282	92.100
	DistilBERT	*97.436*	97.493	94.030	94.018	89.241	88.620	*92.953*	92.778
18-L ResNet-CNN	BERT	94.872	94.868	92.537	93.050	90.419	89.731	92.953	*92.930*
	RoBERTa	97.436	97.500	95.522	95.539	89.241	88.620	92.617	92.459
	DistilBERT	88.889	88.573	97.015	97.101	87.425	86.483	92.282	92.100

Abbreviations: A:Accuracy, F1: F1 score

high-performance on both short and long-text datasets and outperformed the current state-of-the-art models.

The experimental results show that the performance of proposed models improved significantly when trained and tested on larger amount of data which was achieved by combining all the considered datasets. It was also observed that there was no significant difference between different LMs, whether large LM or a light LM. The performance differ by 1-2% only between different LMs. Hence, it can be concluded that combining datasets helps to evaluate the generalised performance of the proposed model.

The results also indicate that a sophisticated model does not guarantee a better outcome. Many studies implied that using models with large number of trainable parameters improve performance. However, a simple 4-layers MLP classifier with simple LM can perform better than combination of complex classifiers (ResNet-CNN) and LMs (BERT or DistillBERT). The difference of results obtained from RoBERTa, BERT and DistilBERT with different classifiers were insignificant: only 1% to 3% difference. Therefore, it is suggested that in real-world scenarios training time, cost and computational complexity of the models should be considered carefully rather than small improvements in resuls. Future studies should examine the interaction between datasets, algorithm models, and splitting of data to understand the solution in more comprehensive manner.

Statements and Declarations

- **Ethics approval and Consent to participate**
  Not applicable
- **Human and Animal Ethics**
  Not applicable
- **Consent for publication**
  Not applicable
- **Availability of data and materials**

Not applicable

- **Competing interests** We confident to confirm that we do not have any conflicts of interest associated with this publication.
- **Funding**
This study is funded by Sebelas Maret University Indonesia.
- **Authors' contributions**
All of the authors contributed equally to this work. The first draft of the article was written by Rini Anggrainingsih, All authors read, commented all the versions of the manuscript, and approved the final manuscript.
- **Acknowledgements**
The authors express gratitude to the Sebelas Maret University for exceptional support as a sponsor of this research.
- **Authors information**
Rini Anggrainingsih, received her bachelor degree and masters from Diponegoro University and Gadjahmada University, respectively. She is currently working as academic staff at Sebelas Maret University while pursuing a PhD at The University of Western Australia. Her current research interest are false information detection on the online platforms, and other natural language processing tasks to improve info-surveillance.
Dr. Ghulam Mubashar Hassan, received his B.Sc. Electrical and Electronics Engineering degree (with honors) from University of Engineering and Technology (UET) Peshawar, Pakistan. He completed his MS in Electrical and Computer Engineering from Oklahoma State University USA and his PhD from The University of Western Australia (UWA) in Joint Schools of Computer Science & Software Engineering and Civil & Resource Engineering. He was valedictorian and received many awards for his PhD. Currently, he is working in UWA and previously he worked in UET Peshawar and King Saud University. His research interests are multidisciplinary problems which include using artificial intelligence, machine learning, pattern recognition, optimization in different fields of engineering and education.
Amitava Datta (A'04–M'10), received the M.Tech. and Ph.D. degrees from IIT Madras in 1988 and 1992, respectively. He did the post-doctoral research at the Max Planck Institut für Informatik, Germany, and the University of Freiburg, Germany. He joined the University of New England in 1995 and The University of Western Australia in 1998, where he is currently a Professor with the School of Computer Science and Software Engineering. His current research interests are in optical computing, data mining, bioinformatics, and social network analysis. He has authored over 150 papers in various international journals and conference proceedings, including the IEEE Transactions on Computers, the IEEE Transactions on Parallel and Distributed Systems, the IEEE Transactions on Visualization and Computer Graphics, the IEEE Transactions on Mobile Computing, the IEEE/ACM Transactions on Networking, the IEEE Transactions on Computational Social Systems and the IEEE Transactions on Systems, Man, and Cybernetics.

References

1. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011, 675–684 (2011). https://doi.org/10.1145/1963405.1963500

2. Ito, J., Song, J., Toda, H., Koike, Y., Oyama, S.: Assessment of tweet credibility with lda features. In: Proceedings of the 24th International Conference on World Wide Web, pp. 953–958 (2015)

3. Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE **11**(3) (2016) arXiv:1511.07487. https://doi.org/10.1371/journal.pone.0150989

4. Hassan, N.Y., Haggag, M.H.: Supervised Learning Approach for Twitter Credibility Detection. 2018 13th International Conference on Computer Engineering and Systems (ICCES), 196–201 (2018)

5. Herzallah, W., Faris, H., Adwan, O.: Feature engineering for detecting spammers on twitter: Modelling and analysis. Journal of Information Science **44**(2), 230–247 (2018)

6. Ahmed, H., Traore, I., Saad, S.: Detecting opinion spams and fake news using text classification. Security and Privacy **1**(1), 9 (2018)

7. Dong, X., Victor, U., Chowdhury, S., Qian, L.: Deep two-path semi-supervised learning for fake news detection. arXiv preprint arXiv:1906.05659 (2019)

8. Reis, J.C., Correia, A., Murai, F., Veloso, A., Benevenuto, F.: Supervised learning for fake news detection. IEEE Intelligent Systems **34**(2), 76–81 (2019)

9. Al-Ahmad, B., Al-Zoubi, A., Abu Khurma, R., Aljarah, I.: An evolutionary fake news detection method for covid-19 pandemic information. Symmetry **13**(6), 1091 (2021)

10. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.-f., Cha, M.: Detecting Rumors from Microblogs with Recurrent Neural Networks. Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), 3818–3824 (2015)

11. Ruchansky, N.: CSI : A Hybrid Deep Model for Fake News Detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 797–806 (2017)

12. Alkhodair, S.A., Ding, S.H.H., Fung, B.C.M.: Detecting breaking news rumors of emerging topics in social media. Information Processing and Management **57**(2), 102018 (2020). https://doi.org/10.1016/j.ipm.2019.02.016

13. Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T.: A Convolutional Approach for Misinformation Identification. IJCAI International Joint Conference on Artificial Intelligence **0**, 3901–3907 (2017). https://doi.org/10.24963/ijcai.2017/545

14. Xu, F., Sheng, V.S., Wang, M.: Knowledge-Based Systems Near real-time topic-driven rumor detection in source microblogs. Knowledge-Based Systems **207**, 106391 (2020). https://doi.org/10.1016/j.knosys.2020.106391

15. Santhoshkumar, S., Babu, L.D.: Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks. Social Network Analysis and Mining **10**(1), 1–17 (2020)

16. Bharti, M., Jindal, H.: Automatic Rumour Detection Model on Social Media. 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), 367–371 (2021). https://doi.org/10.1109/pdgc50313.2020.9315738

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

18. Ajao, O.: Fake News Identification on Twitter with Hybrid CNN and RNN Models. Proceedings of the 9th international conference on social media and society, 226–230 (2018)

19. Kochkina, E., Liakata, M., Zubiaga, A.: All-in-one: Multi-task learning for rumour verification. arXiv preprint arXiv:1806.03713 (2018)

20. Karimi, H., Roy, P., Saba-Sadiya, S., Tang, J.: Multi-source multi-class fake news detection. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1546–1557 (2018)

21. Das, S.D., Basak, A., Dutta, S.: A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles. Neurocomputing (2021)

22. Zhang, P., Ran, H., Jia, C., Li, X., Han, X.: A lightweight propagation path aggregating network with neural topic model for rumor detection. Neurocomputing **458**, 468–477 (2021)

23. Zhou, K., Li, B.: Early Rumour Detection. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, Volume 1, 1614–1623 (2019)

24. Ma, T., Zhou, H., Tian, Y., Al-Nabhan, N.: A novel rumor detection algorithm based on entity recognition, sentence reconfiguration, and ordinary differential equation network. Neurocomputing **447**, 224–234 (2021)

25. Yuan, C., Ma, Q., Zhou, W., Han, J., Hu, S.: Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In: 2019 IEEE International Conference on Data Mining (ICDM), pp. 796–805 (2019). IEEE

26. Wu, Z., Pi, D., Chen, J., Xie, M., Cao, J.: Rumor detection based on propagation graph neural network with attention mechanism. Expert systems with applications **158**, 113595 (2020)

27. Lu, Y.-J., Li, C.-T.: Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648 (2020)

28. Wang, Y., Qian, S., Hu, J., Fang, Q., Xu, C.: Fake news detection via knowledge-driven multimodal graph convolutional networks. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 540–547 (2020)

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

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

31. Moradi, M., Dorffner, G., Samwald, M.: Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. Computer methods and programs in biomedicine **184**, 105117 (2020)

32. Ma, T., Pan, Q., Rong, H., Qian, Y., Tian, Y., Al-Nabhan, N.: T-bertsum: Topic-aware text summarization based on bert. IEEE Transactions on Computational Social Systems (2021)

33. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345 (2019)

34. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)

35. Chen, Y.-C., Gan, Z., Cheng, Y., Liu, J., Liu, J.: Distilling knowledge learned in bert for text generation. arXiv preprint arXiv:1911.03829 (2019)

36. Qu, Y., Liu, P., Song, W., Liu, L., Cheng, M.: A text generation and prediction system: Pre-training on new corpora using bert and gpt-2. In: 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 323–326 (2020). IEEE

37. Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., Li, W.: The automatic text classification method based on bert and feature union. In: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), pp. 774–777 (2019). IEEE

38. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics, pp. 194–206 (2019). Springer

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

40. Rietzler, A., Stabinger, S., Opitz, P., Engl, S.: Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. arXiv preprint arXiv:1908.11860 (2019)

41. Yu, J., Jiang, J.: Adapting bert for target-oriented multimodal sentiment classification. In: IJCAI,2019, pp. 323–326 (2019). IJCAI

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

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

44. Koirala, A.: Covid-19 fake news dataset. In: Mendeley Data, p. 1 (2021). Mendeley. 10.17632/zwfdmp5syg.1

45. Bondielli, A., Marcelloni, F.: A survey on fake news and rumour detection techniques. Information Sciences **497**, 38–55 (2019). https://doi.org/10.1016/j.ins.2019.05.035

46. Meel, P., Vishwakarma, D.K.: Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. Expert Systems with Applications **153**, 112986 (2020)

47. Pamungkas, E.W., Basile, V., Patti, V.: Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure. arXiv preprint arXiv:1901.01911 (2019)

48. Ghenai, A., Mejova, Y.: Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017, 518 (2017) arXiv:1707.03778. https://doi.org/10.1109/ICHI.2017.58

49. Chatterjee, S., Deng, S., Liu, J., Shan, R., Jiao, W.: Classifying facts and opinions in twitter messages: a deep learning-based approach. Journal of Business Analytics **1**(1), 29–39 (2018)

50. Wu, L., Rao, Y.: Adaptive interaction fusion networks for fake news detection. arXiv preprint arXiv:2004.10009 (2020)

51. Wu, L., Rao, Y., Zhao, Y., Liang, H., Nazir, A.: Dtca: Decision tree-based co-attention networks for explainable claim verification. arXiv preprint arXiv:2004.13455 (2020)

52. Wang, Z., Guo, Y.: Rumor events detection enhanced by encoding sentimental information into time series division and word representations. Neurocomputing **397**, 224–243 (2020)

53. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)

54. Anggrainingsih, R., Hassan, G.M., Datta, A.: Bert based classification system for detecting rumours on twitter. arXiv preprint arXiv:2109.02975 (2021)

55. Kaliyar, R.K., Goswami, A., Narang, P.: Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia tools and applications **80**(8), 11765–11788 (2021)

56. Gupta, P., Gandhi, S., Chakravarthi, B.R.: Leveraging transfer learning techniques-bert, roberta, albert and distilbert for fake review detection. In: Forum for Information Retrieval Evaluation, pp. 75–82 (2021)

57. Huang, Y.-H., Liu, T.-W., Lee, S.-R., Calderon Alvarado, F.H., Chen, Y.-S.: Conquering cross-source failure for news credibility: Learning generalizable representations beyond content embedding. In: Proceedings of The Web Conference 2020, pp. 774–784 (2020)

58. Pelrine, K., Danovitch, J., Rabbany, R.: The surprising performance of

simple baselines for misinformation detection. In: Proceedings of the Web Conference 2021, pp. 3432–3441 (2021)

59. Taud, H., Mas, J.F.: Multilayer Perceptron (MLP) Neural networks. Geomatic Approaches for Modeling Land Change Scenarios, 451–455 (2018)

60. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

61. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)

62. Ma, J., Gao, W., Wong, K.-F.: Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In: The World Wide Web Conference, pp. 3049–3055 (2019)

63. Dong, X., Victor, U., Chowdhury, S., Qian, L., View, P.: Deep two-path semi-supervised learning for fake news detection. arXiv preprint arXiv:1906.05659 (2019) arXiv:arXiv:1906.05659v1