UNIVERSITY OF PERADENIYA

FACULTY OF SCIENCE

CONTEXTUALIZED FAKE NEWS DETECTION USING TRANSFORMER MODELS

DECLARATION

We hereby declare that the Project Summary Report entitled **("Title of the project"**) is an authentic record of our own work as a requirement of the three-months project under the course of '**Independent study in Data Science (DSC3263)**' during the period from 01/11/2023 to 26/02/2024 for the award of the degree of B.Sc. Honours Study in Data Science from the Department of Statistics and Computer Science Faculty of Science University of Peradeniya, under the guidance of (Dr. Sachith P. Abeysundara, Prof. Roshan D. Yapa).

P. K. K. N. S. Jayathilake	W. G. A. C. Chandrasena	A.P.S.I. De Waas Gunawardhana
S/18/406	S/18/332	S/18/337

Date: 19/02/2024

Certified by:

1. **Supervisor (Name) :** Prof. Roshan D. Yapa

   Date: ……………………………..

   (Signature):………………………..

2. **Head of the Department (Name):** Dr. Sachith P. Abeysundara

   Date : ……………………………..

   (Signature):………………………..

Department Stamp:

ABSTRACT

The rapid development of the Internet allows a quick spread of information through social networks or websites. Without concern about the credibility of the information, unverified or fake news is spread on social networks and reaches thousands of users. Fake news is typically generated for commercial and political interests to mislead and attract readers. The spread of fake news has raised a significant challenge to society. It is the reason the 21st century is known as the post-truth era. This paper demonstrates a model and methodology for fake news detection. The main objective is to build a transformer model that can be used to differentiate contextualized "Real" news and "Fake" news. We use a fine-tuned base Bidirectional Encoder Representation from Transformers (BERT) model and a traditional Recurrent Neural Network (RNN) model, the Bidirectional Long Short-Term Memory (BiLSTM) model on a publically available dataset for the detection of fake and real news. Bert model performs 97.85% accuracy on test data, whereas the BiLSTM model performs 93.65% accuracy. Our result shows that the BERT model improves 4.2% of the accuracy in fake news detection compared to the BiLSTM Model. Since the base BERT model has a maximum token size, we used news with a maximum of 150 words. As a future study, this model can be improved to differentiate news with a large word count.

ACKNOWLEDGEMENTS

We wish to extend our deepest gratitude to Prof. Roshan D. Yapa, Senior Professor in the Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya and Mr. Buddhika Kasthuriarachchy, Head of Data Analytics, N-able (Pvt) Ltd, Sri Lanka for their valuable guidance and supervision to the completion of our project.

Our heartfelt gratitude to Dr.Ruwan D. Nawarathna, Dr.Hakeem A. Usoof, Dr. Ruwanthini Siyambalapitiya, Dr.Mahasen Dehideniya & Ms. B.R.Pvithra M. Basnayake for their valuable insights, guidance, technical directions, knowledge sharing and encouragement to the completion of our project.

We are grateful to Mr. Mapa Wakkumbura & Mr. Dinesh Deepal for assisting us on our technology requirement in Department.

TABLE OF THE CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 01

1.1. Error! Bookmark not defined.

Fake news has become a pressing concern in an era dominated by information. Rumors, misinformation, disinformation, and malformation are common challenges confronting media of all types. It is, however, worse in the case of digital media, especially on social media platforms. Social media is one of the major platforms to get news and information. However, it also provides convenience for widespread fake news. The reason behind fake news is to create hype to get the audience's attention and build a negative impact on society. The spread of false and misleading news has led to significant social and economic consequences, impacting industries from finance to healthcare. Misinformation and fake news can have a long-term impact, mainly when people rely on accurate information to make critical decisions. The need to detect fake news has never been more crucial. (Muhammed T & Mathew, 2022)

The term "fake news" refers to disinformation or false information presented as legitimate news. It can be challenging to manually identify and counteract fake news due to its sheer volume and rapid dissemination. Machine-learning and deep-learning approaches have demonstrated accurate predictions and insights to handle complicated problems (Jain et al., 2019). Developing automatic, trustworthy, and accurate solutions for detecting fake news is a hot research area. Detecting fake news is a challenging Natural Language Processing (NLP) problem concerned with text classification to distinguish between fake and real. NLP has advanced significantly over the past few years. Transformer-based pre-trained language models are now the state-of-the-art approach for many NLP problems. However, studying fake news detection using transformer-based models is still limited. (Qazi et al., 2020)

In this paper, we use a transformer model as our primary model. It is a fine-tuned base BERT model. We use a traditional RNN model, the BiLSTM model, to compare the accuracy of BERT model. Our main objective is to build a transformer model to that can use to differentiate contextualized “Real” news and “Fake” news.

The paper is structured as follows. Chapter 01, Introduction of Fake News Detection, briefly overviews prior research works and the literature review. Chapter 02 describes the research methodology, which includes details of the models. Chapter 03 discusses the analysis, findings, results, and discussion, including comparison of two models. Chapter 04 concludes with the future works and limitations of the research.

Keywords: *BERT, Fake News Detection, Neural Network, LSTM, Transformer model*

1.2. Literature Review

1.2.1. LSTM

Long Short Term Memory Networks (LSTM) are a recurrent neural network type also known as RNN. (Bahad et al., 2019) LSTM networks are capable of learning long-term dependencies in sequential data, which makes them well-suited for tasks such as language translation, speech recognition, and time series forecasting. The input gate controls which data from the present moment's input is permitted into those memory cells. It is made up of a layer that is sigmoid and a method of point-wise multiplication that keeps the values between 0 and 1. If the sign of the output is 1, that input is completely accepted, however a value of 0 means the input is completely banned. (Amity School of Engineering and Technology Lucknow, Amity University Uttar Pradesh, India & Singh, 2023)

The LSTM architecture consists of a set of recurrently connected sub-networks, known as memory blocks. The idea behind the memory block is to maintain its state over time and regulate the information flow thought nonlinear gating units. Fig. 1 displays the architecture of a vanilla LSTM block, which involves the gates, the input signal x (t), the output y (t), the activation functions, and peephole connections. The output of the block is recurrently connected back to the block input and all of the gates. (Van Houdt et al., 2020)

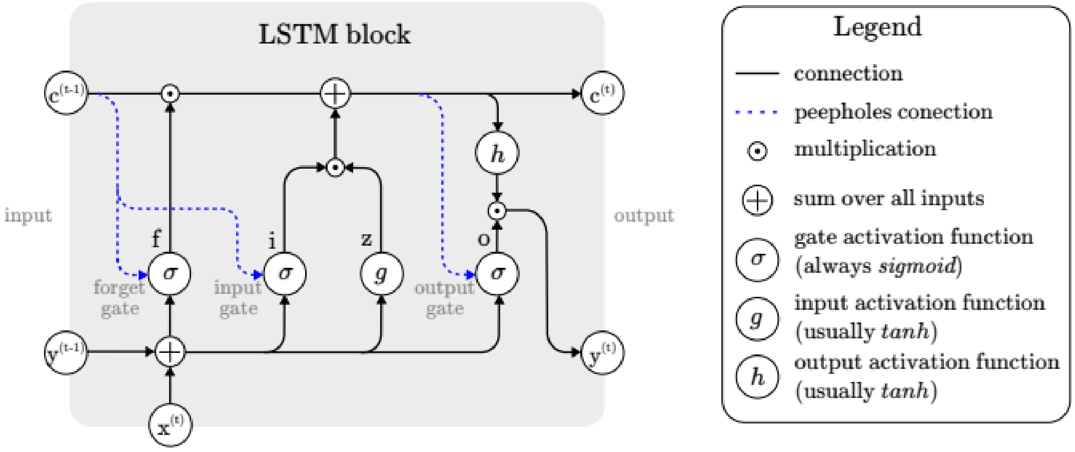


Figure 1. 1: Architecture of a typical Vanila LSTM block.

Long Short-Term Memory networks (LSTM) are a special type of RNN competent in learning long-term dependencies. LSTM is a very effective solution for addressing the vanishing gradient problem. In LSTM-RNN the hidden layer of basic RNN is replaced by an LSTM cell as in Figure.

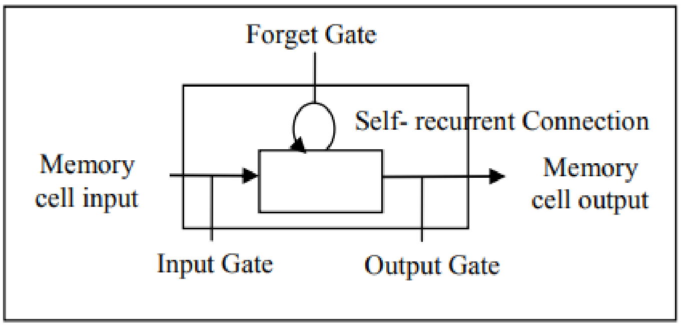


Figure 1. 2: Structure of LSTM cell

1.2.2. Bidirectional LSTM

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. In problems where all time steps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.

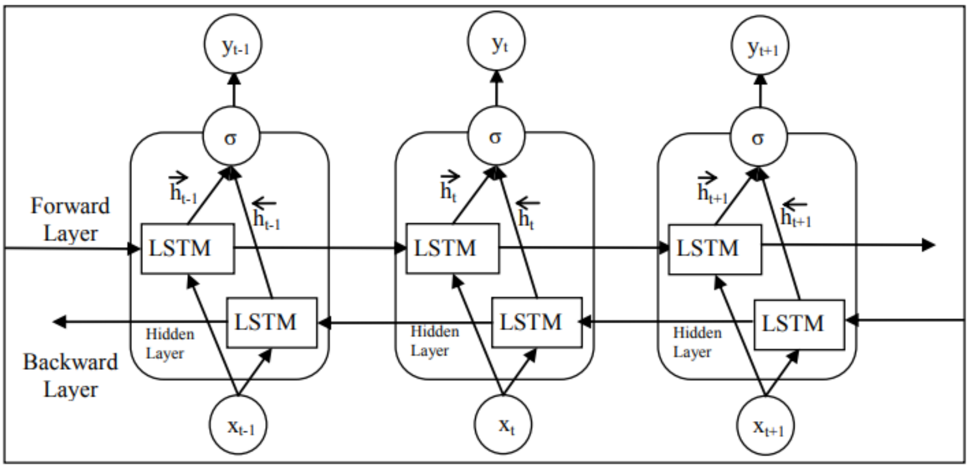


Figure 1. 3: General Architecture of Bi-directional LSTM-RNN

1.2.3. Transformer Models

The best performing models also connect the encoder and decoder through an attention mechanism. We(Vaswani et al., 2023) propose a new simple network architecture, the Transformer, based solely on attention mechanisms. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. (Vaswani et al., 2023) Transformer is a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output.

The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in the left and right halves of Figure 1.4, respectively.

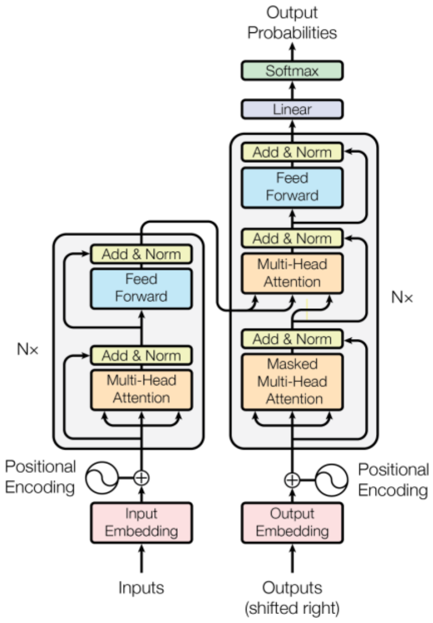


Figure 1. 4: The Transformer - Model Architecture

1.2.4. BERT

BERT (Devlin et al. 2018), initially developed by Google, has its origins from pretraining contextual representations learning in NLP. It can handle NLP tasks such as supervised text classification, question-answering, text summarization, without the need for human intervention. BERT performs pre-training using an unsupervised prediction task, which includes a masked language model (MLM) and a next sentence predictor. MLM is about understanding context first and then predicting words. The input is included in the Transformer structure to predict the masked words based on the context of the surrounding words. Through these processes, BERT understands the context more accurately. The next sentence predictor is for identifying the relationship between sentences. This task is important for language understanding tasks such as Question Answering (QA) or Natural Language Inference (NLI).(Jwa et al., 2019)
This model structure allows BERT to perform very well in various NLP tasks. (Jwa et al., 2019; Vaswani et al., n.d.)

1.2.5. Model Architecture

BERT is based on a multi-layer bidirectional transformer encoder that jointly conditions both the left and the right contexts in all layers.(Jwa et al., 2019)

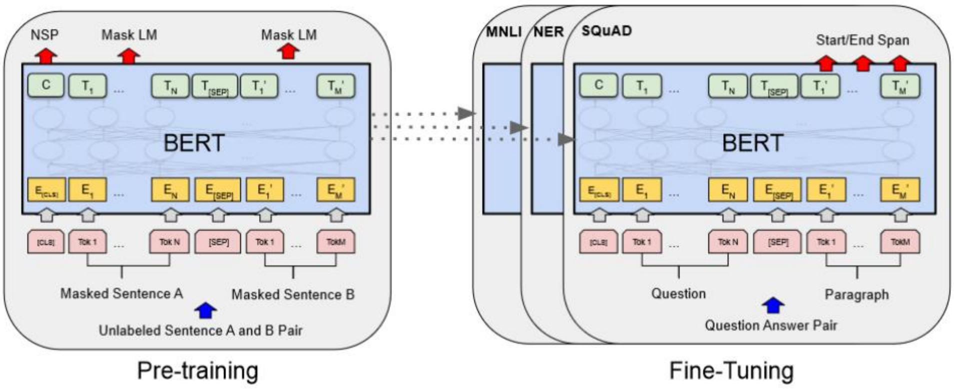


Figure 1. 5: Pre-training and fine-tuning procedures for BERT

Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers) (Devlin et al., 2019). The BERT model employs an attention mechanism to acquire the contextual associations among words in the input text of the transformer. (Devlin et al., 2019; Koru & Uluyol, 2024) BERT uses the Transformer architecture, which has Attention as its main goal.(Azizah & Widiarto, n.d.).

The attention equation,

$$\text{attention (Q, K, V)} = \text{SoftMax } (\frac{\mathrm{Q}K^T}{\sqrt{d_k}})\, \boldsymbol{V}$$

where:

- Q is matrix that constructs the query (contains a vector of each word),
- K is key,
- V is value itself.
- dk is dimension of key vector K.
- $\frac{\mathrm{Q}K^T}{\sqrt{d_k}}$ is the step for calculating attention weight, which is the result of the dot product between Q and K divided by the square root of d_k.
- SoftMax = one of the activations in deep learning that is useful for attention weights that lead to between 0 and 1 (a kind of probability).

1.3. Related Work

1.3.1. Related to Sri Lanka

- In Sri Lanka, University of Vavuniya conducted a fake news detection study regarding covid-19 which is an analysis of Machine Learning Algorithms. This study has implemented two extraction features and compared them with eight different machine learning techniques such as Support Vector Machine (SVM), logistic regression, decision tree, etc.

1.3.2. Related to other countries

- This section briefly summarises the work in the field of fake news detection. Following is the work that had an impact on our findings.
- (Kumar & Shah, 2018) have explored a comprehensive survey of diverse aspects of fake news. Different categories of fake news, existing algorithms for counterfeit news detection, and future aspects have been explored in this research article.
- (Koru & Uluyol, 2024) Is based on Detection of Turkish Fake News from Tweets with BERT Models. They fine-tuned a pre-trained BERT deep learning model, and extended variations of the model with Bi-LSTM and Convolutional Neural Network (CNN) layers with the frozen and unfrozen parameters methods were explored.
- (Sr & Ahmad, 2024) have presented a new approach for detecting fake news from news posted on social media. They proposed using a probabilistic fusion strategy to combine the knowledge gained from two language models BERT-CNN and BERT-LSTM, at a classification score level. In this paper they mentioned that Under varying parameter settings, the detection accuracy attained supersede the existing fake news detection methods by at least 3%.
- In one of the researches, (Joy et al., 2022) the authors have conducted a comparative analysis by implementing five transformer-based models such as BERT, BERT without LSTM, ALBERT, RoBERTa, and a Hybrid of BERT & ALBERT in order to detect the fraudulent news of COVID-19 from the internet. According to their results, the RoBERTa model has performed better than other models by obtaining an F1 score of 0.98 in both real and fake classes.
- Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection was conducted in Indonesia (Azizah & Widiarto, n.d.) In that research, they explore those transformer models and found that ALBERT outperformed other models with 87.6% accuracy, 86.9% precision, 86.9% F1-score, and 174.5 run-time (s/epoch) respectively.
- In another study, the accuracy of Bi-directional LSTM-RNN model with CNN, vanilla RNN, and unidirectional LSTM-RNN are evaluated and compared.(Bahad et al., 2019). From this study they observed that CNN performs better for extracting local and position-invariant features while LSTM-RNN is well suited for a long-range semantic

dependency based classification. Further, the results show that Bi-directional LSTM-RNN model is significantly more effective than unidirectional models.

- Another paper discusses LSTM (Long Short-Term Memory), Bidirectional LSTM (BiLSTM), and Convolution Neural Network (CNN)-based algorithms for identifying fake news.(Amity School of Engineering and Technology Lucknow, Amity University Uttar Pradesh, India & Singh, 2023) . Author stated that, in terms of both accuracy and F1-score, the CNN beat the standard LSTM and BiLSTM models. CNN-BiLSTM is the most effective model.

CHAPTER 2

2.1. Error! Bookmark not defined.

2.1.1. Primary data

The Fake News Corpus, an open-source dataset primarily made up of millions of news articles collected from a carefully curated list of 1001 sites from http://www.opensources.co/ (currently,the website is broken), serves as the main dataset used for the false news identification challenge. In order to better balance the classes, articles from the NYTimes and WebHose English News Articles have been added, as the list does not include many trustworthy sources. The corpus was produced by scraping every domain that http://www.opensources.co/ provided (using scrapy python library). The newspaper library was then used to parse all of the raw HTML material and extract the article text along with a few more fields mentioned below.

Table 2. 1: Overview of the features for the Fake News Corpus

id	Unique number as the index
domain	Domain link of the web link to the article
type	Given in the table 2 below
url	Web link to the article
content	Content of the article
scraped_at	Scraped date and time
inserted_at	Uploaded date and time
updated_at	Final updated date and time
title	Title of the article
authors	Authors mentioned in the article
keywords	Keywords extracted from the title and the content

meta_keywords	Another type of keywords extracted from the content
meta_description	Small description about the content
tags	Tags for the content
summary	Summary extracted from the content column
source	Open sources, NYTimes or a WebHose English News article

The primary objective of the corpus is to train deep learning algorithms for the identification of fake news. According to the dataset description it contains merely 9,408,908 articles (745 out of 1001 domains) are present in the public edition. The label assigned to each article matches the label linked to its domain. There are 10 distinct 1GB sized zip files containing the news corpus. The WinRAR application was used to extract the entire dataset in comma separated values (csv) format after downloading each zip file. Then the dataset was uploaded into a Jupyter notebook as two distinct Pandas data frames after using only the content, title, and type columns because the completely extracted csv file is around 27 GB in size.

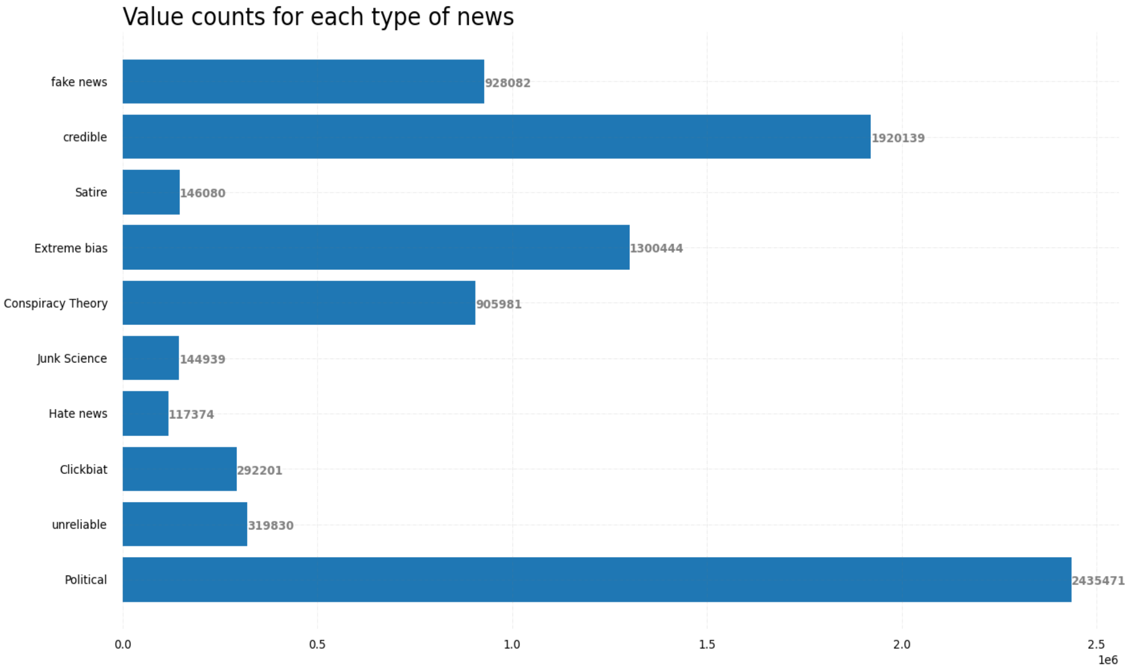


Figure 2. 1: Graph of the news types in 'type' column (according to the website)

Dataset : Fake News Corpus
All the source codes about the dataset are available at FakeNewsRecognition

2.1.2. Data preparation

pre-processing steps had to be taken to ensure that the data would be fit for the task at hand. The first step was to discard the null values related to the content, title, and type columns. Since the dataset in very large, removing null values does not significantly impact the number of data in the dataset. There was no duplicate data with respect to the type, title, and content columns. After removing the null values, a new data frame was created using the types fake and reliable. The total number of fake and reliable data were 2,848,222. In order to prevent biased results as a consequence of class-imbalance (Jacobusse & Veenman, 2016) sampling was performed to select a random subset of fake data equal to the total number of reliable data. Thus, the dataset was balanced with 50,000 instances of each label class. Then a new column 'label' was appended to the data set by replacing the words 'fake' as 0 and 'reliable' as 1 from the type column. From here onwards the new dataset is called as the sample dataset, consists of 100,000 records with label, title and content columns.

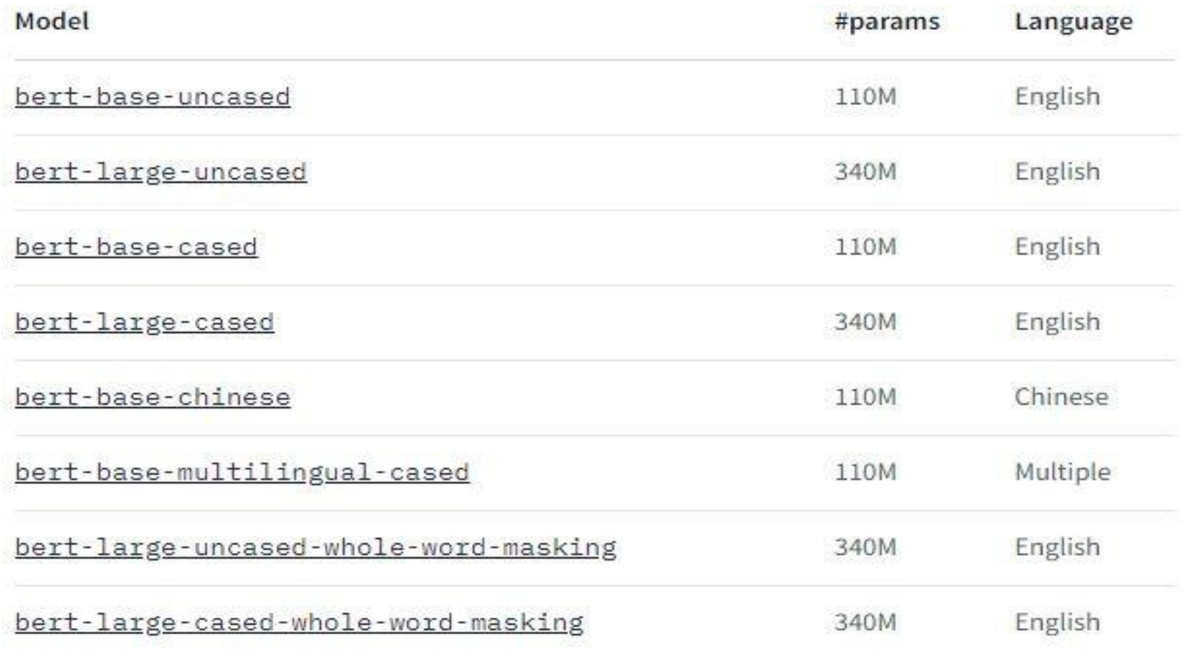
2.1.3. Text data pre-processing

Traditionally, the initial stage in natural language processing (NLP) tasks is text pre-processing, also known as text normalisation (Sprung et al., 2010). Text processing can perform noticeably better when the right pre-processing techniques are chosen, taking into account the domain and language of the textual input. After displaying some randomly selected titles and contents, in both content and title columns, there were web links (urls), email addresses, digits, new line symbol (\n), and unwanted white spaces. Web links, email addresses, and digits were replaced with suitable tags. New line characters and the extra white spaces were replaced with a single white space.

Table 2. 2: Unwanted features and the replacement tags

Fact	**Replacement tag**
Web links (https:// and www.)	url
Email addresses	email
Digits	0

2.1.4. BERT models



Model	#params	Language
bert-base-uncased	110M	English
bert-large-uncased	340M	English
bert-base-cased	110M	English
bert-large-cased	340M	English
bert-base-chinese	110M	Chinese
bert-base-multilingual-cased	110M	Multiple
bert-large-uncased-whole-word-masking	340M	English
bert-large-cased-whole-word-masking	340M	English

Figure 2. 2: Different types of basic BERT models

Source: https://huggingface.co/google-bert/bert-base-uncased

For this project, we use "bert-base-uncased" because it is smaller in size, computationally affordable, and not applicable to complex text mining operations. The following table shows some differences between the BERT base and BERT large models.

Table 2. 3: Some of main differences between BERT base and BERT large

Model name	**Encoder layers**	**Hidden units**	**Attention heads**
BERT base	12	768	12
BERT large	24	1024	16

2.1.5. BERT Tokenizer

A tokenizer is in charge of preparing the inputs for a model. Here we used BertTokenizer from the pre-trained bert-base-uncased model. The main tasks of a tokenizer as follow.

- Tokenizing (splitting strings in sub-word token strings), converting tokens strings to ids and back, and encoding/decoding (i.e., tokenizing and converting to integers).
- Adding new tokens to the vocabulary in a way that is independent of the underlying structure.
- Managing special tokens (like mask, beginning-of-sentence, etc.): adding them, assigning them to attributes in the tokenizer for easy access and making sure they are not split during tokenization.

BERT is pre-trained on a fixed vocabulary learned from the corpora that the model is trained on. BertTokenizer is a tokenizer that is based on the WordPiece model. From the corpora that BERT is trained on, this tokenizer establishes a vocabulary of 30,000 tokens from the data. In addition to words, the tokenizer also established tokens for all unique single characters as well as sub-words (Devlin et al., 2019). For each entry in the vocabulary, a vocabulary ID is also assigned. This ID is used to link the tokens to their respective word embedding vectors. In addition to splitting input words into sub-units, all input sentences are prepended with a special classify token ([CLS]) and appended with a separate token ([SEP]). For multi-sentence inputs, BERT only prepends a single [CLS] token. Sub words are prepended with two “#” characters. An example of the pre-processing conducted by BERT is as follows:

Table 2. 4: Example of Pre-Processed text by BERT Tokenizer

Original input	Let's try to tokenize!
Tokenized data	['let', "'", 's', 'try', 'to', 'token', '##ize', '!']
Word embaddings (input ids)	{'input_ids': [101, 2292, 1005, 1055, 3046, 2000, 19204, 4697, 999, 102], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]}

In here input ids 101 and 102 represents the special tokens [CLS] and [SEP] respectively.

2.1.6. Input Data for the BERT model

The maximum number of input tokens for the BERT model is 512, which means the number of input tokens must not exceed 512. Because of that, we had to filter out a maximum number of words in the input paragraph for the content column. After several attempts, we found that the safest maximum number of words is around 150. By filtering out 10,000 fake and real data values from the sample dataset, we prepared a new dataset as an input dataset with 20,000 records with label, cleaned title, and cleaned content columns.

2.1.7. Configuration of the BERT model

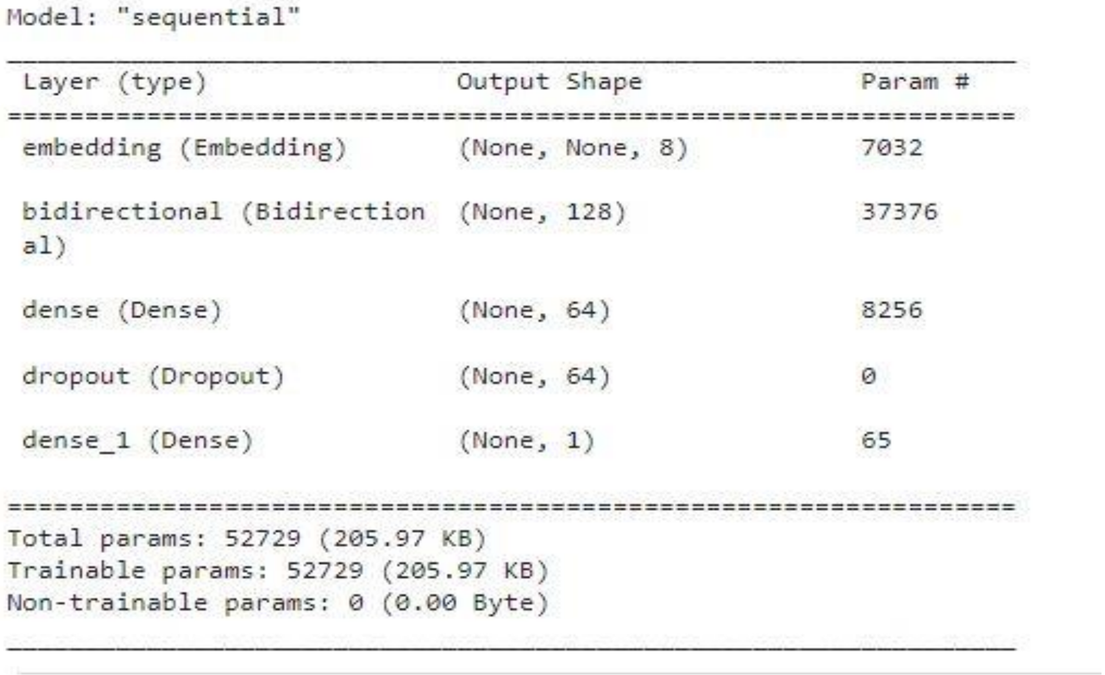
Table 2. 5: Configuration of the BERT model used for training

Parameter	**Description**	**Value**
Vocab_size	Vocabulary size of the BERT model. Defines the different tokens that can be represented by the input_ids passed to the forward method of BERT model	30522
hidden_size	Dimensionality of the encoder layers and the pooler layer.	768
num_hidden_layers	Number of hidden layers in the Transformer model	12

num_attention_heads	Number of attention heads for each attention layer in the Transformer encoder.	12
intermediate_size	Dimensionality of the “intermediate” (i.e., feed-forward) layer in the Transformer encoder.	3072
hidden_act	The non-linear activation function (function or string) in the encoder and pooler. If string, “gelu”, “relu”, “swish” and “gelu_new” are supported (Zhang et al., 2021).	gelu
hidden_dropout_prob	The dropout probabilitiy for all fully connected layers in the embeddings, encoder, and pooler.	0.1
attention_probs_dropout_prob	The dropout ratio for the attention probabilities.	0.2
max_position_embeddings	The maximum sequence length that this model might ever be used with. Typically set this to something large just in case	512
type_vocab_size	The vocabulary size of the token_type_ids passed into BERT model	2
initializer_range	The standard deviation of the truncated_normal_initializer for initializing all weight matrices.	0.02
layer_norm_eps	The epsilon used by the layer normalization layers.	1e-12
gradient_checkpointing	If True, use gradient checkpointing to save memory at the expense of slower backward pass.	False

2.1.8. BiLSTM Model

In this project Bi-LSTM is used in order compare the transformer model BERT with traditional RNN (Kaliyar et al., 2021). The same dataset which used in the BERT was used as the input dataset in BiLSTM.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 8)           7032

 bidirectional (Bidirection  (None, 128)               37376
 al)

 dense (Dense)               (None, 64)                8256

 dropout (Dropout)           (None, 64)                0

 dense_1 (Dense)             (None, 1)                 65

=================================================================
Total params: 52729 (205.97 KB)
Trainable params: 52729 (205.97 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 2. 3: Bi-LSTM model configuration

2.1.9. Model hyperparameters

Table 2. 6: List of hyperparameters for BERT and Bi-LSTM

Hyperparameter	**BERT**	**Bi-LSTM**
Batch size	32	64
Epoch	10	10
Optimizer	Adam	Adam
Loss function	Binary crossentropy	Binary crossentropy
Dropout rate	0.1	0.2
Learning rate	1e-5	1e-3

2.1.10. Training, validation and testing datasets

In order to do the comparisons, for both models we used the same training, validation and testing splitting. Out of 20,000 records, the number of records belongs to each dataset were given below.

Table 2. 7: Sizes of the training, validation and testing datasets

Dataset	**Size**
Training dataset	16,000
Validation dataset	2,000
Testing dataset	2,000

2.1.11. Evaluation

In this study, we used accuracy, precision, recall and f1-score to compare the two models:

$$Accuracy = \frac{Number\ of\ correct\ predications}{Total\ number\ of\ predictions}$$

$$precision = \frac{TruePositve}{TruePositive + FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive + FlaseNegative}$$

$$F1 = 2.\frac{Precision.Recall}{Precision + Recall}$$

CHAPTER 03

3.1. Findings

This section briefly overviews some descriptive analysis for ***clean_content*** column in the final data set we used to train our models. (After replacing web links, emails and digits with appropriate tags and without removing stop words).

Table 3. 1: Analysis of cleaned content column in the final data set

Total No of Words	9008750
Total Number of Unique Words	879
Maximum No of Words	150
Mean No of Words	75

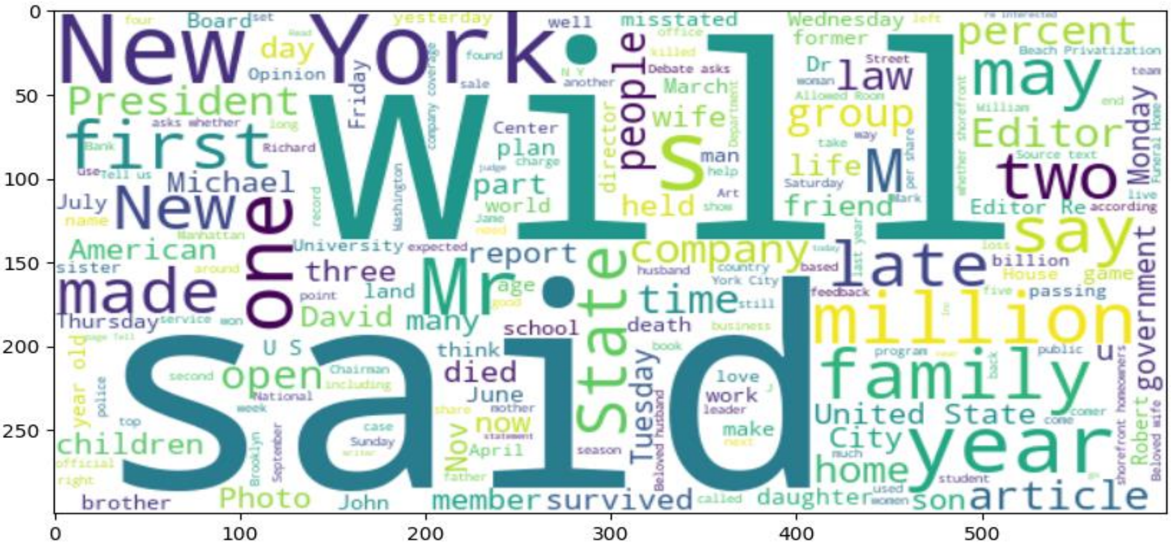


Figure 3. 1: Word cloud of reliable news for most frequent 200 words

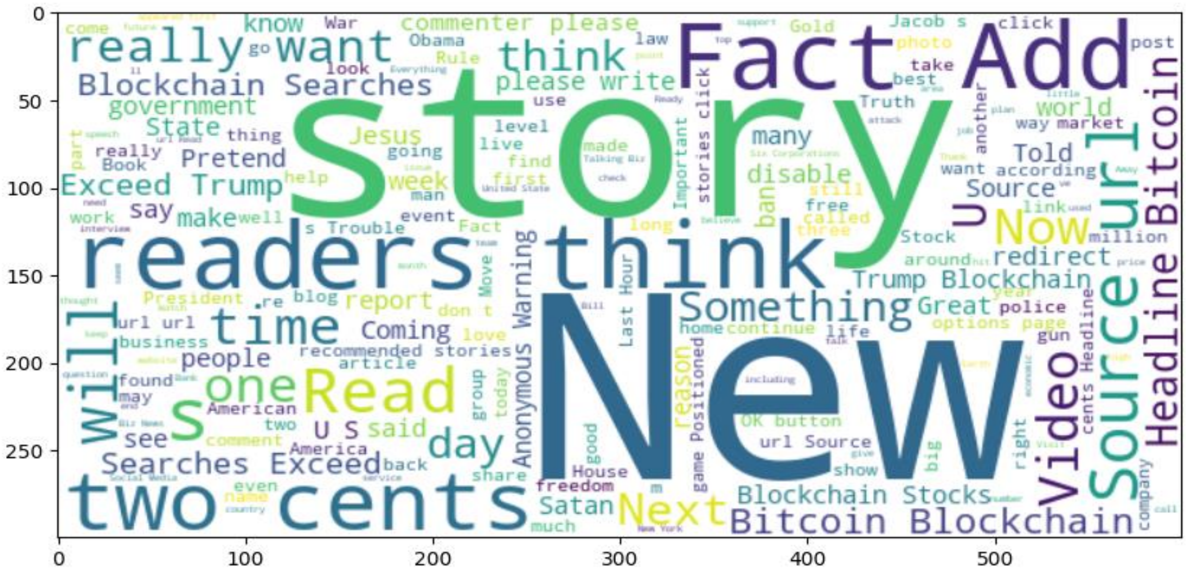


Figure 3. 2: Word cloud of fake news for most frequent 200 words

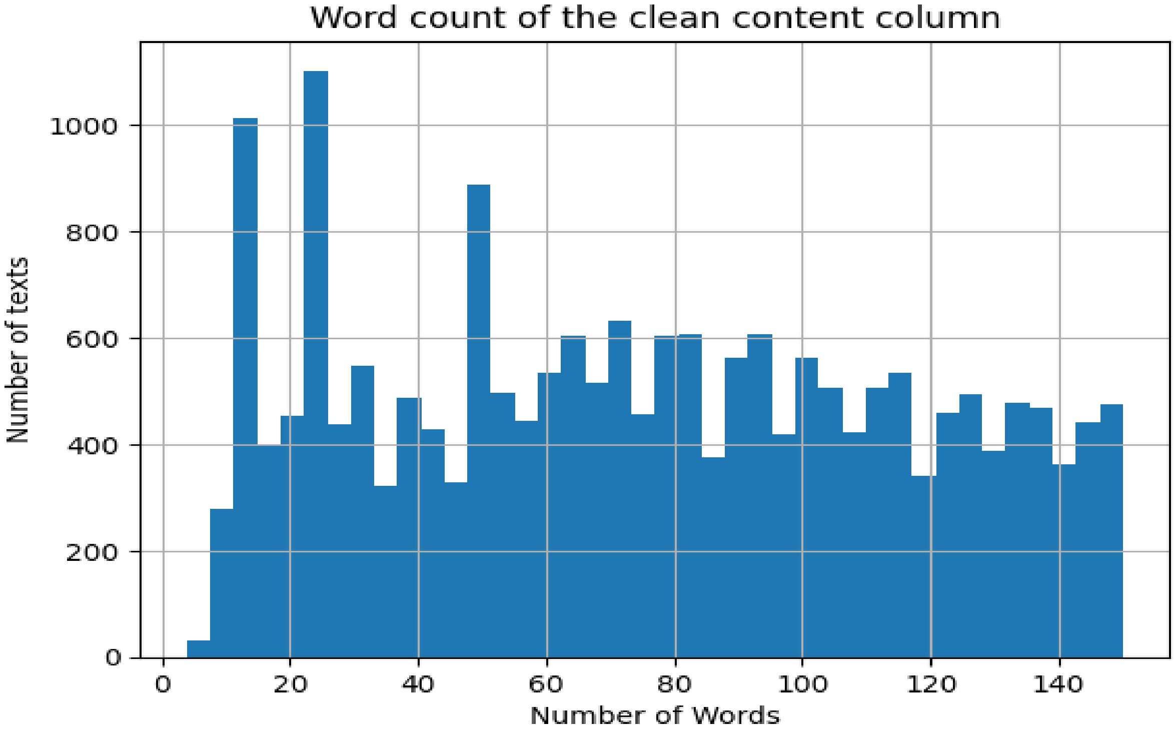


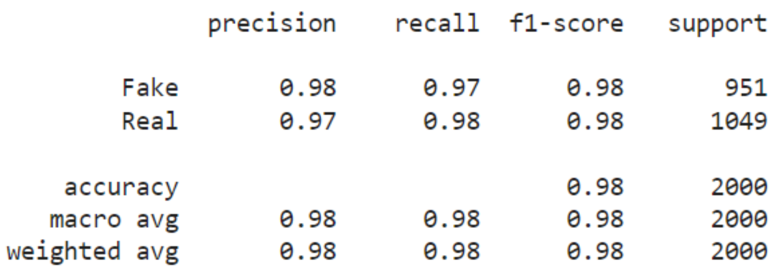
Figure 3. 3: Histogram of word count of the clean content column after pre-processing

3.2. Results

In this section shows the results of each model and compared the performance of BERT model and BiLSTM model using several evaluation metrics like accuracy, precision, recall, weighted f1-score. The results of the various experiments on the test set are reported in Table 3.1.2.1.

The results clearly showing that Transformer based model, BERT model is considerably better than BiLSTM, RNN model for our Contextualized Fake News Detection task.

3.2.1. Results of the Model 1 : BERT model



```
              precision    recall  f1-score   support

        Fake       0.98      0.97      0.98       951
        Real       0.97      0.98      0.98      1049

    accuracy                           0.98      2000
   macro avg       0.98      0.98      0.98      2000
weighted avg       0.98      0.98      0.98      2000
```

Figure 3. 4: Classification Report of BERT Model

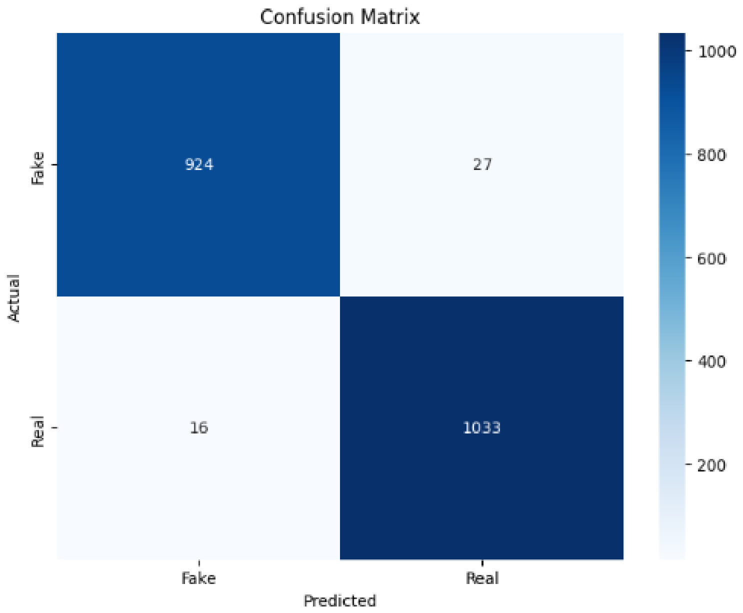


Figure 3. 5: Confusion Matrix of BERT Model

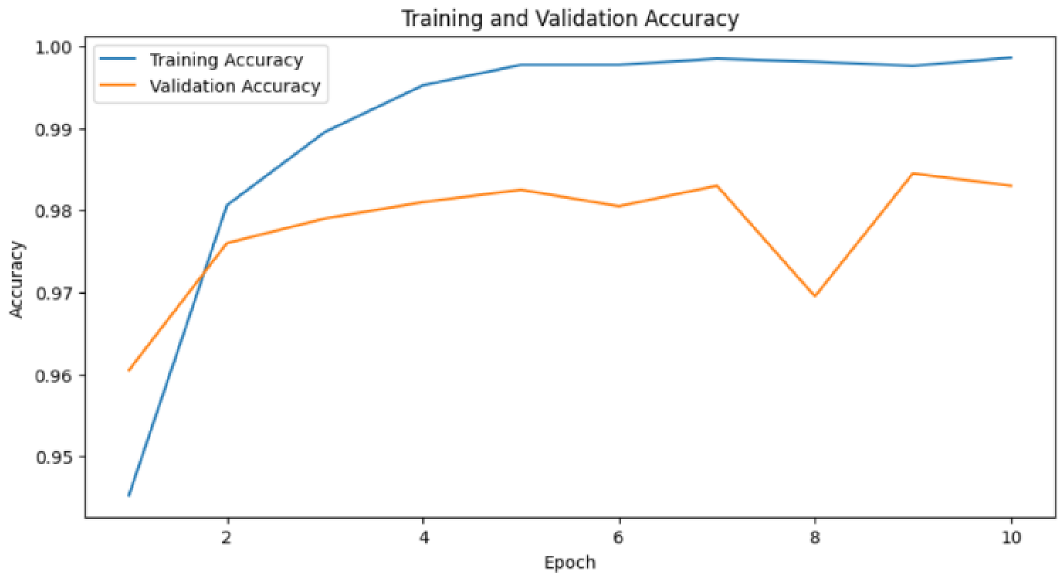


Figure 3. 6: Training and Validation Loss Curve for BERT model

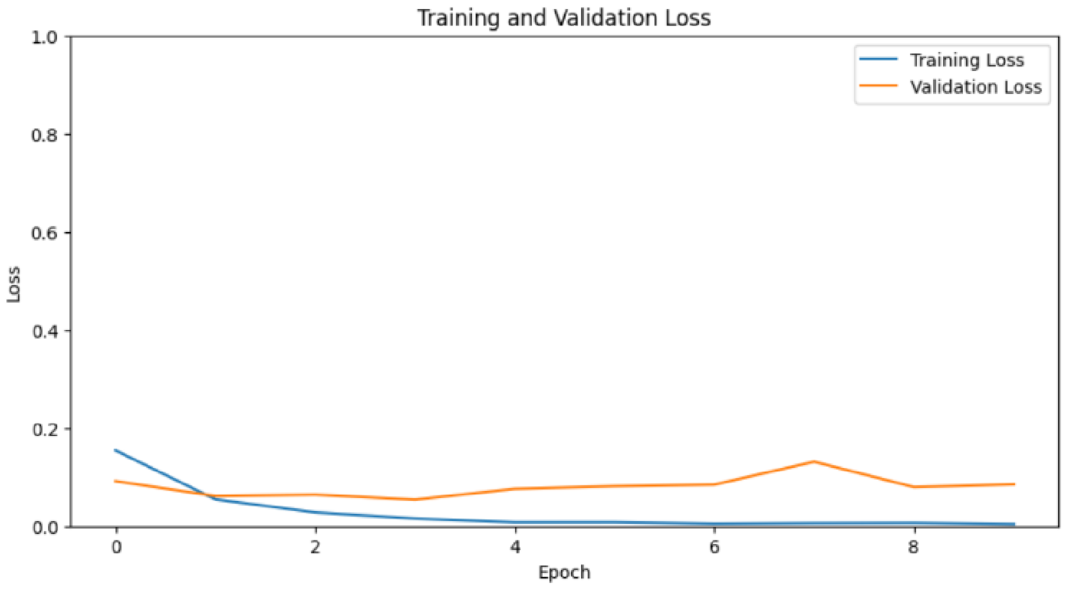
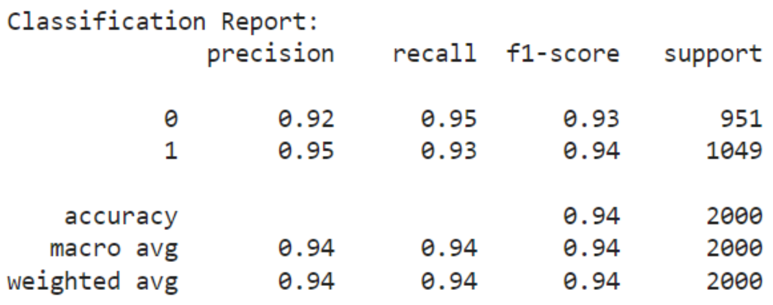


Figure 3. 7: Training and Validation Accuracy Curve for BERT Model



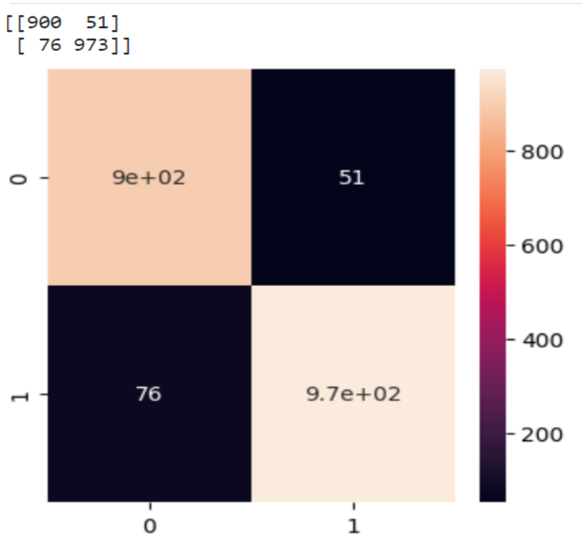
```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.95      0.93       951
           1       0.95      0.93      0.94      1049

    accuracy                           0.94      2000
   macro avg       0.94      0.94      0.94      2000
weighted avg       0.94      0.94      0.94      2000
```

Figure 3. 8: Classification Report for BERT Model

3.2.2. Results of the Model 2 : LSTM model



```
[[900  51]
 [ 76 973]]
```

Figure 3. 9: Confusion Matrix of BiLSTM Model

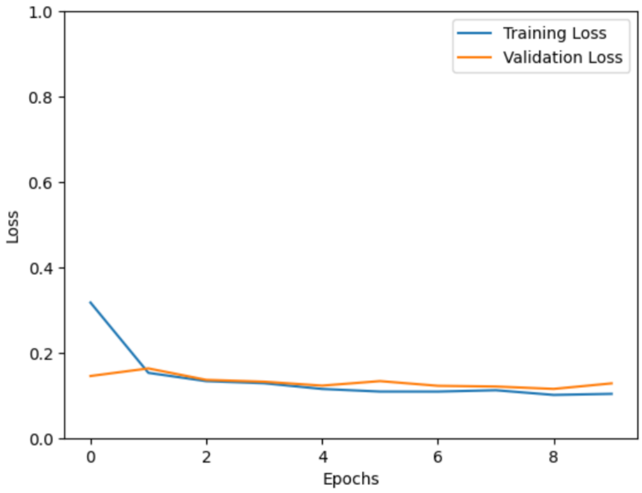


Figure 3. 11: Training and Validation Loss for BiLSTM Model

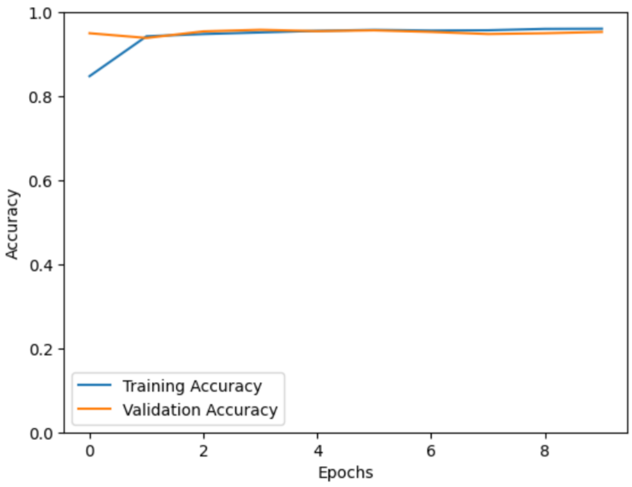


Figure 3. 10: Training and Validation Accuracy for BiLSTM Model

3.3. Comparison of BERT Model & BiLSTM Model

Table 3. 2: Comparison of various fake news detection models on final data

Model Type	Model	Accuracy	Precision	Recall	F1-Score
Transformer Model	BERT	0.9785	0.9745	0.9847	0.9796
RNN Model	BiLSTM	0.9365	0.9502	0.9276	0.9388

3.4. Discussion**Error! Bookmark not defined.**

We trained our BERT model in the Department Computer Workstation. Training took 4-5 days on NVIDIA Qudro RTX 4000(8GB) GPU. Experimental environment was setup on Google Colabory for BiLSTM model. Python scripts used for executing the experiments discussed in this are available in a GitHub repository. The link to the GitHub repository is https://github.com/Fake_News_Detection_Model.
The result shows that the BERT model improves 4.2% of the accuracy in fake news detection compared to the BiLSTM Model. In the BiLSTM model, we did not remove any stop words to ensure that both models use the same input data. Though the Accuracy and Loss graphs of the BERT model are not very accurate, increasing the number of epochs and further fine-tuning can get accurate graphs. Since model training takes a considerable amount of time, even for 10 epochs, we did not train this BERT model with many epochs.

CHAPTER 04

4.1. Limitations

BERT model has a maximum token size (512) we had to limit the maximum word count of a news to a number less than 512. After several try and fail attempts we found safest maximum word count of a news is 150. Since we are interested on implement this model to identify Sri Lanka's news either in Sinhala or English we searched for a data set. Though there were several research carried there was not quality data set to do that. (Jayawickrama et al., n.d.) This model requires a tabular formatted data set contain *content* and *label* columns. Also, when we train our models we used maximum epochs size of 20 with our resources. With sufficient resources we can go further numb epochs.

4.2. Conclusion

The paper presents a fine-tuned BERT model to improve fake or real news detection. A publicly available data set is used for detection purposes. Fake news is a progressively significant and tricky problem to solve. The literature study shows that various ML-based detection procedures are introduced to detect fake news. However, these models need higher performance. We improve performance by using an attention mechanism-based transformer model. The result shows that the proposed methodology of the fine-tuned BERT model improves the accuracy of detection over the traditional BiLSTM model. The proposed model works well for the balanced news data set where a particular news has words at most 150. However, Human judgment over the text cannot be detected easily. Therefore, we need further improvement in detection techniques.

4.3. Future works

- As we mentioned in limitation section, since the base BERT model has a maximum token size, we used news with a maximum of 150 words. As a future study, this model can be improved to differentiate news with a large word count. Applying BERT to long text classification tasks will need more computational power. For instance, a 1,500-token text needs about 14.6GB memory to run BERT-large even with batch size of 1, exceeding the capacity of common GPUs (e.g. 11GB for RTX 2080ti). Moreover, the O(L^2) space complexity implies a fast increase with the text length L (Ding et al., n.d.).
- Combine several columns such as title and author to the content column, in order to get more accurate results for future predictions. This will also help to investigate the authors and organizations, who are publishing fake and reliable news articles.
- Use BERT summarization functions such as bert-extractive-summarizer to summarize paragraphs having huge number of words into a smaller number of words. This will allow to control the token sizes and reduce the computational power as well(Abdel-Salam & Rafea, 2022)
- Fine-tune RoBERTa transformer model to identify reliable and fake news posted in Sinhala language.

REFERENCES

1. Abdel-Salam, S., & Rafea, A. (2022). Performance Study on Extractive Text Summarization Using BERT Models. *Information*, *13*(2), 67. https://doi.org/10.3390/info13020067
2. Amity School of Engineering and Technology Lucknow, Amity University Uttar Pradesh, India, & Singh, Y. (2023). Fake News Detection Using LSTM in TensorFlow and Deep Learning. *Journal of Applied Science and Education (JASE)*, *3*(2), 1–14. https://doi.org/10.54060/jase.v3i2.35
3. Azizah, S. F. N., & Widiarto, W. (n.d.). *Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection*.
4. Bahad, P., Saxena, P., & Kamal, R. (2019). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *Procedia Computer Science*, *165*, 74–82. https://doi.org/10.1016/j.procs.2020.01.072
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805
6. Ding, M., Yang, H., Zhou, C., & Tang, J. (n.d.). *CogLTX: Applying BERT to Long Texts*.

7. Jacobusse, G., & Veenman, C. (2016). On Selection Bias with Imbalanced Classes. In T. Calders, M. Ceci, & D. Malerba (Eds.), *Discovery Science* (Vol. 9956, pp. 325–340). Springer International Publishing. https://doi.org/10.1007/978-3-319-46307-0_21

8. Jayawickrama, V., Ranasinghe, A., Attanayake, D. C., & Wijeratne, Y. (n.d.). *A Corpus and Machine Learning Models for Fake News Classification in Sinhala*.

9. Joy, S. K. S., Dofadar, D. F., Khan, R. H., Ahmed, Md. S., & Rahman, R. (2022). A Comparative Study on COVID-19 Fake News Detection Using Different Transformer Based Models. *2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, 1–5. https://doi.org/10.1109/ISIEA54517.2022.9873797

10. Jwa, H., Oh, D., Park, K., Kang, J., & Lim, H. (2019). exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences*, *9*(19), 4062. https://doi.org/10.3390/app9194062

11. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

12. Koru, G. K., & Uluyol, Ç. (2024). Detection of Turkish Fake News From Tweets With BERT Models. *IEEE Access*, *12*, 14918–14931. https://doi.org/10.1109/ACCESS.2024.3354165

13. Kumar, S., & Shah, N. (2018). *False Information on Web and Social Media: A Survey* (arXiv:1804.08559). arXiv. http://arxiv.org/abs/1804.08559

14. Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. *International Journal of Data Science and Analytics*, *13*(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6

15. Qazi, M., Khan, M. U. S., & Ali, M. (2020). Detection of Fake News Using Transformer Model. *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 1–6. https://doi.org/10.1109/iCoMET48670.2020.9074071

16. Sprung, C. L., Cohen, R., & Adini, B. (2010). Chapter 1. Introduction. *Intensive Care Medicine*, *36*(S1), 4–10. https://doi.org/10.1007/s00134-010-1760-5

17. Sr, S. M., & Ahmad, S. (2024). BERT based Blended approach for Fake News Detection. *Journal of Big Data and Artificial Intelligence*, *2*(1). https://doi.org/10.54116/jbdai.v2i1.27

18. Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, *53*(8), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (n.d.). *Attention Is All You Need*.

21. Zhang, X., Chang, D., Qi, W., & Zhan, Z. (2021). A Study on Different Functionalities and Performances among Different Activation Functions across Different ANNs for Image Classification. *Journal of Physics: Conference Series*, *1732*(1), 012026. https://doi.org/10.1088/1742-6596/1732/1/012026