RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt25613

Deceptive News Content Detection using a Hybrid Transformer-based and Deep Learning Model with Explainability

Arati M Chabukswar [a,*], P. Deepa Shenoy [a], S.M. Dasharath [b], K.R. Venugopal [a]

[a] Department of Computer Science & Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, 560001, Karnataka, India

[b] School of Mechanical Engineering, REVA University, Bengaluru, 560064, Karnataka, India

* Corresponding Author Email: arati.chabukswar0106@gmail.com

DOI: https://doi.org/10.54392/irjmt25613

Received: 28-07-2025; Revised: 15-11-2025; Accepted: 22-11-2025; Published: 28-11-2025

Abstract: The growth of social media platforms has facilitated knowledge dissemination. The ability of misinformation to affect elections, public opinion, and instigate instability makes it a dangerous threat to civilization that spreads rapidly. For an informed and reliable information ecology to survive, the ability to identify deceptive information in an extensive variety of languages is essential. The Transformer based pretrained language models (TB-PLMs) like Distilled Bidirectional Encoder Representations from Transformers (DistilBERT), ALite BERT (ALBERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa) versions of the BERT model with a deep neural network structures such as Bi-directional Gated Recurrent Unit (BiGRU) and Convolution Neural Network (CNN) is used for the identification of deceptive news in English. The dataset utilized for the challenge consists of a combination of LIAR, and Fake/Real news dataset, resulting in a Combined Corpus (CC) dataset about politics. TB-PLMs are optimized to understand the semantic linkages and contextual information found in the dataset. BiGRU and CNN layers are used to capture the dependencies between neighboring characters in the text. The experimental findings show that the RoBERTa+BiGRU model performs better in comparison with all the other models in identifying English deceptive news with an accuracy of 99.04%. The results obtained reflect that RoBERTa+BiGRU has rise of 0.06% in accuracy from the base RoBERTa model. Also, the results of proposed work on DistilBERT+BiGRU and RoBERTa+BiGRU model performs well based on features (class 0 and class 1) while utilizing Local Interpretable Model-agnostic Explanations (LIME) implementation to clarify the target labels which can facilitate valid data extraction and processing to successfully counteract deceptive information.

Keywords: Deep learning, Fake News, Natural Language Processing, Political News, Transformer Based Pre-Trained Language model (TB-PLMs), XAI – LIME

1. Introduction

The propagation of deceptive information epidemic has significantly expanded over the past ten years, helped along by social media. There are several reasons to disseminate this falsified news. Certain ones are created only with the intention of boosting website traffic and hits to influence public opinion regarding financial markets or political issues harming the organizations and businesses online reputations. Social media fake news about health poses a threat to world health. In February 2020, the World Health Organization issued a warning that the COVID-19 pandemic had been accompanied by a large "infodemic" or an excess of information [1]. Digital channels, including websites, online forums, and social media, have surpassed traditional media as the primary information sources in the modern day [2].

The authors of [3] talk on the spread of deceptive information by Indian political parties like the Bhartiya Janata Party and Indian National Congress about issues like gender, celebrities, religion, and so forth. Both people and society may suffer because of fake news. Over the years, several studies on automatic deceptive news identification have been conducted employing transformer-based pre-trained models, traditional machine learning, and deep learning techniques [4-7]. Nonetheless, most of them concentrated on finding news on COVID-19 and political news. As a result, they created their models and characteristics for datasets that corresponded with their area of interest. Nevertheless, the comparative research that has already been done on deceptive news detection techniques has also concentrated on a certain kind of dataset or examined a small number of models. For instance, Wang created the benchmark dataset LIAR

and used it to test a few pre-existing models [4]. Research on sophisticated pretrained language model like BERT with CNN and exBAKE an extra unlabeled news corpus into BAKE that enhances BERT for the identification of deceptive news has been done in [8, 9] respectively.

Transformer-based systems can encode deeper semantic and contextualized information about a particular input text, whereas DL approaches enable the capture of more prominent and important information. To leverage both, the model is suggested with multiple designs built upon the pre-trained transformer model that employ various neural-based structures such as BERT with Bi-LSTM and Bi-GRU layers with WELFake dataset [10].

To identify political fake news, this study investigates the potential of many pretrained transformer models, including DistilBERT, ALBERT, RoBERTa [11]. To investigate the effectiveness of neural network, structures added to DistilBERT, ALBERT, RoBERTa designs for improving political fake news detection, few neural network structures are added on top of DistilBERT, ALBERT, RoBERTa (fine-tuning techniques).

The proposed study has shown that several pre-trained transformer-based models exhibit varying levels of performance over a range of techniques. For instance, it has been demonstrated that, when combined with a BiGRU layer, the pretrained RoBERTa performs optimally when compared to all other suggested techniques. In general, promising results have been observed when pretrained transformer models are extended utilizing powerful DL models like CNN and BiGRU. The primary contributions of the proposed work can be summed up as follows:

- Based on few downstream neural network topologies, unique transferring transformer-based approaches are investigated to achieve optimal performance.
- Examining how the fine-tuning technique affects the suggested transferring methodologies to strengthen the contention that the strategy has a strong chance of enhancing performance even more.
- A comparison and analysis of the performance differences between the widely used models and the suggested models are made after extensive tests utilizing pretrained-transformer based models and DL models by utilizing the datasets from existing work in the suggested model's working environment.
- Also, the work is carried out on usage of XAI (Explainable AI) LIME for better understanding on determining the important factors affecting the classification.

The format of this paper is as follows. In Section 2, previous work on the subject is reviewed. The proposed approach is covered in Section 3. The ideas for results, discussion and insights are presented in Section 4. Section 5 includes Comparative Analysis of proposed with existing work and Section 6 concludes the paper.

2. Related Work

This section provides a summary of the relevant literature on political news, with a focus on transformer-based methodologies that leverage the CombinedCorpus dataset, which is created by merging the “LIAR” and “Fake and Real news” datasets. It draws attention to a selection of earlier research that employs sophisticated and traditional machine learning-based techniques for detecting deceptive news. Diverse methodologies, including transformer-based models, deep learning techniques, and conventional machine learning (ML), have been used to identify deceptive information. Table 1 is providing the information on research that are frequently conducted in the path of Deceptive News detection (DND) on social media.

2.1 Conventional and advanced machine learning (ML) models

Using ML and DL classifiers in a Chrome environment, the authors in [12] suggests an autonomous method for Facebook's fake news identification. Using an LSTM algorithm, a Chrome extension was designed to assess user profiles and news article features with 99.4% accuracy. To overcome the limitations of earlier research, the authors in [13] suggests an ensemble classification approach to increase the accuracy of deceptive news identification. This model extracts and classifies characteristics from deceptive news datasets using Random Forest, Decision Tree, and Extra Tree Classifier. In [14], authors have proposed a method called CreditRank for calculating credibility of publishers on social networks, a framework for early detection of fake news using user and content-based features.

2.2 Advanced transfer learning language model detection

The study [15] focus on employing highly accurate pre-trained BERT and RoBERTa models that have been fine-tuned using actual and fake COVID-19 news to identify and stop the spread of deceptive information. BERT model performs better than RoBERTa. The research [16] investigates the automated analysis of political utterances in Romanian using cutting-edge natural language processing algorithms and assesses how crucial context is for confirming their accuracy.

Table 1. Research that is frequently carried out in deceptive information detection

Ref	Datasets	Contribution	Accuracy (%)	Challenges	Future Enhancement
[6]	NELA-GT-19 and Fakeddit.	FND-NS model (news content and social contexts) proposed. BART combined with rich features into encoder, used as sequence-to-sequence transformer along with BERT and GPT-2.	FND-NS model: 74.8%	Addressed early fake news detection, label shortage.	Work with NELA-GT20 or scrape more news sources. Use semi-supervised learning. Expand experiments with better infrastructure.
[18]	COVID-19 tweets	Finetuning BERT, CTBERT models, Exploring ML model impacts. Performance evaluation of CNN and BiGRU models.	CTBERT+BiGRU: 98.46%	Obtained dataset from Constraint@AAAI 2021 COVID-19 fake news detection by participating in competition.	Explore diverse hyperparameters and optimization methods. Work with larger datasets.
[19]	LIAR, Fake or real news, Combined corpus.	Performance analysis on 19 ML methods using 3 datasets; 8 ML models, 6 DL models, 5 transformer-based models (BERT, RoBERTa, DistilBERT, ELECTRA, ELMo).	RoBERTa accuracy on combined corpus - 96%, on Fake or Real dataset - 98%.	Building a combined corpus of 80k news including a wide range of topics.	Designing few more models that can detect misinformation and health-related fake news.
[20]	LIAR, Fake and Real News	Usage of “Llama” Large language model proposed by Meta with 7 billion parameter version.	Not mentioned	High training time, GPU memory requirement.	Apply adapters for 13B, 30B, and larger models for parameter-efficient fine-tuning.
[21]	Twitter, Facebook, Instagram COVID-19 fake news dataset	Usage of BERT (base, large), ALBERT (base, large, xlarge), RoBERTa (base, large), CT-BERT, Ro-CT-BERT for FND.	Ro-CT-BERT 99. 0187%	Short news sentences with uncommon professional terms, heated up softmax loss to identify hard-mining samples.	Apply adapters for 13B, 30B, and larger models for parameter-efficient fine-tuning.
[22]	FakeNewsNet, KaiDMML	Document level embedding -BERT. Sentence level embedding – BERT, RoBERTa, DistilGPT2, xlNet, DistilBERT.	Not mentioned	More training time required to run the model for 15 epochs on KaiDMML dataset.	Use larger databases, scrape content from public websites.
[23]	LIAR, ISOT, COVID-19	Usage of single and multilayer perceptron’s, CNN after embedding layer consisting of BERT, RoBERTa, GPT-2, and Funnel Transformer.	Funnel-CNN: 99.96%	More training time required to run the model for 15 epochs on the KaiDMML dataset.	Usage of larger datasets.

A fresh dataset from the Factual project, whose results, when compared to PolitiFact studies, provide a solid baseline utilizing the RoBERTa model.

Using a variety of DL models, the study [17] investigates automated methods for identifying fake news regarding COVID-19 on social media, with a peak accuracy of 98.41% on the Constaint@AAAI 2021 Covid-19 Fake News Detection dataset. It assesses various supervised text classification techniques, such as CNN, LSTM, and BERT, emphasizing the need of pre-training language models using unlabeled COVID-19 tweets. Transformer-based models perform better than other models, with BERT in particular showing because it is beneficial to pre-train on a domain-specific corpus before fine-tuning. Table 1 shows the research that is frequently carried out in deceptive information detection with challenges faced while performing the task.

A thorough analysis of deep learning and machine learning techniques for identifying fake news is presented in [24]. They examine methods, difficulties, and developments in 90 recent investigations. Transformer-based models perform better than standard machine learning, although they have problems with interpretability, generalizability, and dataset quality. To enhance performance and reliability, they suggest more study on explainable, lightweight, and multilingual models. The intrinsic features of FND such as intentional manufacture, heteromorphic transmission, and controversial reception are examined in the article [25] providing a thorough grasp of the topic by summarizing current detection techniques based on these features with future fields by discussing technological developments. In [26], authors propose a hybrid DistilBERT + BiLSTM model for fake news detection that combines high accuracy with interpretability using LIME to identify significant textual aspects, exceeding baseline transformer approaches while giving transparent, human-understandable insights.

Authors in [27] integrates text and image data using BERT and cross attention to detect fake news. For the domain-agnostic multilingual deceptive news classification, authors in [28] suggests an efficient neural model based on the multilingual BERT. Additionally, five languages and seven distinct domains were worked on in a multilingual, multidomain deceptive news detection dataset. To achieve improved accuracy on three datasets, the work in [29] introduces a novel hybrid fake news detection system that combines a light gradient boosting machine (LightGBM) model with BERT-based embeddings. The suggested strategy works better than state-of-the-art methodologies, different word embedding techniques, and classification strategies. Authors uses hybrid architectures BERT (or RoBERTa) with RNN [30] for detection of misinformation on COVID-19 dataset. The study [31] suggests X-FRAME, a hybrid model that combines theory-driven characteristics with XLM-RoBERTa embeddings. Several datasets from the news and social media domains are used to evaluate it. Using LIME and permutation feature significance, the model offers explainability while emphasizing significant elements like sensationalism and source reliability. X-FRAME provides interpretable forecasts and exhibits efficient cross-domain generalization.

The maximum accuracy of 98.39% is attained by RoBERTa with summarisation. assessing the detection of false information produced by GPT-2. Summarisation lowers processing costs and increases accuracy in [32]. In [33], authors demonstrated that pre-trained transformer ensemble models, including XLM-RoBERTa, mBERT, and ELECTRA, to identify deceptive information in a variety of languages. Th work in [34] uses transformers with explainable AI to do reasoning-based explanations on multimodal FND i.e., on text and image. Two major innovations are presented in the study [35], a multilingual multimodal dataset of news articles with paired images various languages including English and HEMT-Fake, a Hybrid Explainable Multimodal Transformer that combines relational, text, and image signals with hierarchical explainability. News text is represented by BERT embeddings and then sequentially modelled using a BiLSTM layer in [36]. For every prediction, it uses SHAP (SHapley Additive Explanations) to produce interpretable feature contributions. The method promotes openness in model decisions by identifying the words or phrases that have the greatest influence on whether COVID-19 news is authentic or fraudulent.

3. Proposed Methodology

This section presents the TB-PLMs such as DistilBERT, ALBERT, and RoBERTa architecture-based deceptive news detection process.

3.1 Fake News Dataset on political news

The proposed study integrates datasets from two sources, namely LIAR (publicly available) and FakeRealNews (FR news) taken from the Kaggle false news dataset, which included articles about the 2016 US presidential election, was made by George McIntire. The combined datasets are referred to as CombinedCorpus (CC) which is used in the proposed work. Table 2 shows the detailed statistics of political news whereas Figure. 1 gives the trend of count of the records on political news datasets. On the online community repository GitHub (https://github.com/JunaedYounusKhan51/FakeNewsDetection), this dataset can be accessed by the public under the FakeNewsDetection dataset. The 50224 text records from two distinct sources—"LIAR" and "Real and Fake news"—are combined to generate the dataset. The suggested work divides the labels into two categories: legitimate (1) and deceptive (0).

Table 2. Number of records on political news datasets

Number of records	Total data
Legitimate (Test+Train)	4782 + 22129 = 26911
Deceptive (Test+Train)	2242 + 21072 = 23314

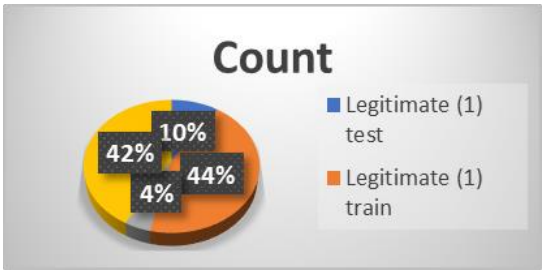

Figure 1 Distribution of number of records on political news datasets

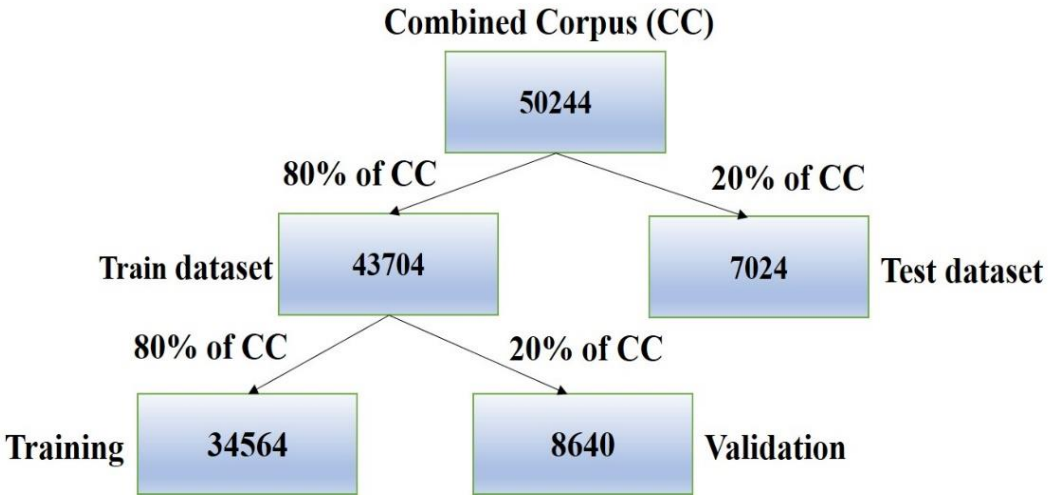

Figure 2. Data splitting into datasets for testing, validation, and training

About 12.8K brief remarks with human classifications having a statement that is 56% true and 44% deceptive are included in the LIAR dataset which has six LIAR labels namely pants-fire, false, barely true, half true, mainly true, and true —are transformed into the binary news classification labels of true and deceptive. Whereas half-true, mostly-true, and true are regarded as legitimate; false, scarcely true, and pants-fire are seen as deceptive. Quotes about political issues from both Democrats and Republicans may be found in most of the posts in this dataset, which come from social media websites.

The actual news is gathered from media sources like Bloomberg, the Wall Street Journal, and the New York Times, while the fake news pertaining to the 2016 US presidential election was gathered for this dataset's misinformation part from the Kaggle fake news dataset [19].The GitHub repository for the dataset has approximately 6.3k news items in total, with half of those being political news. The corpus has an equal distribution of real and deceptive news. As shown in the Figure. 2, the dataset is divided into 80% train and 20% test, further the training dataset is considered as validate dataset with a division proportion of 0.2 for improved performance estimation of model.

3.2 Preprocessing of Text and Pre-trained models

3.2.1 Preprocessing of Text

Political news dataset needs to be pre-processed before being fed into the models. Initially converting the text into lower case, eliminating stop words, replacement of multiple whitespaces with one space, removal of punctuations and special characters. Following that, the spelling errors are cleaned present in the corpus, usage of NLTK toolkit to tokenize the sentences. Every text is divided by white space, and words are lemmatized to eliminate suffices to get its root word lemma. Finally, the cleaned text of words/tokens are rejoined by whitespace. Also, the training data is divided into three categories using samples from the

train dataset and the validation dataset, which have a validation split ratio of 0.2 as train with 34564 records, test with 7024 records and validate datasets with 8640 records.

Additionally, after lemmatizing, the information is sent into the DistilBERT, ALBERT and RoBERTa tokenizer since often the lemmatized words comprise characters that are not exactly consistent with the tokenizer's vocabulary. The text preprocessing steps are explained briefly as follows

- ***Identification of missing values:*** Text components labelled as values that are missing (non-numerical) in the train and test data are replaced with a string that is empty. As a result, mistakes are prevented during model training and a precise representation of the data is ensured.
- ***Conversion of text into Lowercase*:** Each word of the text is changed to lowercase letters using it. Because "Cat" and "cat" are considered synonyms in most tasks involving classification, the model must accumulate few unique attributes.
- ***Removal of Special Symbols and Punctuation:*** Regular expressions are used to remove punctuation and special characters from the text. By doing this, the impact of superfluous characters is reduced, allowing the model to focus on word meaning.
- ***Standardizing Whitespace:*** To ensure consistent tokenization, several consecutive whitespace characters are converted to a single space. This stops the model from misinterpreting the quantity of spaces as a useful characteristic.
- ***Tokenization:*** The NLTK, word tokenize method splits the prepared text into discrete words (tokens). This generates a set of characters that the model can process.
- ***Lemmatization:*** Every token is lemmatized using the lemmatizer. Words are reduced to their most basic form, which improves the approach's universality by encapsulating words' meanings regardless of grammatical modifications.
- ***Stop Word Elimination:*** A list of stop words was used to eliminate certain words from the text. This reduces the dimension of the input and prevents the model from focusing on irrelevant aspects.
- ***Tokens with filters:*** Figure. 3 Illustrates the set of tokens which is left over when stop words are removed from lemmatized tokens.
- ***Merging Written Text:*** At last, the filtered and cleaned tokens are combined once more into a single string. In this fashion, the final pre-processed text representation is generated.

To ensure that the text is further tokenized and normalized in accordance with the model's particular vocabulary and tokenization rules, the DistilBERT, ALBERT and RoBERTa tokenizers are used after lemmatization. The text data is standardized and prepared for effective model training by the above organizing techniques. This improves the quality of the data and allows the model to focus on the most important features for text classification.

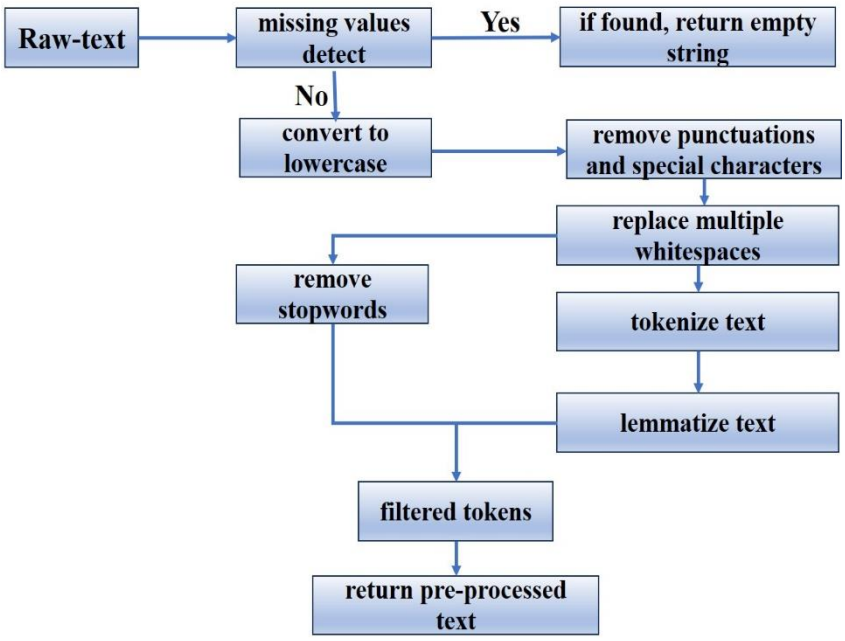

Figure 3. Preprocessing steps for text

3.2.2 Pre-trained model's Text Preprocessing

A DistilBERT tokenizer, an ALBERTTokenizer, and a RoBERTaTokenizer objects are loaded by the pre-trained "DistilBERT-base-cased", "ALBERT-base-v2", and "RoBERTa-base" models. The language and subword components of the DistilBERT, ALBERT, and RoBERTa models are specifically designed to be compatible with this tokenizer.

- ***Encoding and Tokenization of text:*** Using the ALBERT, DistilBERT, and RoBERTa tokenizers, the processed text collection is converted into mathematical representations suitable for the ALBERT, DistilBERT, and RoBERTa models respectively. To protect the model against very long sequences that can jeopardize training uniformity, each tokenizer receives a group of parsed strings together with padding and truncation settings.
- ***Input Features Recovery:*** The input_id and attention mask are the two important models' components that are fetched from the encoders. Input_ids consist of ID numbers for each text model's wrapped token whereas attention mask has padding tokens of value 0 and valid tokens of value 1 in the series of padding which makes the model to ignore padding components and focus on real content.

Using the tokenizers of the pre-trained ALBERT, DistilBERT, and RoBERTa models, the text preprocessing step converts the cleaned text input into a structure suitable for the model's architecture as shown in Figure. 4.

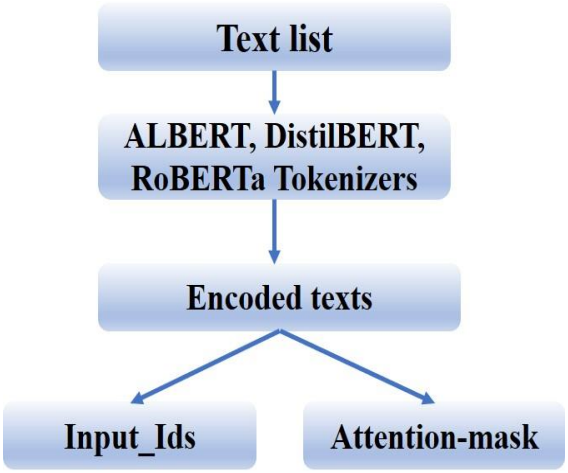

Figure 4 Text pre-processing of pretrained models

3.3 Examined Features

3.3.1 Pretrained features of DistilBERT-base-cased, ALBERT-base-v2, and roberta-base

Contextualized embedded words are used to obtain semantic descriptions for use in the Feature Type. "DistilBERT-base-cased" is a shortened and expedited variant of BERT that maintains 97% of its efficacy while requiring 40% fewer parameters and 60% faster inference. It differentiates between both uppercase and lowercase characters since it is case-sensitive, hence the name cased [11]. "ALBERT-base-v2" evaluates vast volumes of text input while it is being trained, considering a word's meaning in connection to its neighbouring words [5]. This enables it to record semantic details and contextual representations for every word in a sentence. "roberta-base" model which is a robustly optimized BERT method outperforms BERT by using longer sequences and more data for training and the model is uncased, so it cannot distinguish between letters in upper and lowercase [11].

3.3.2. Pretrained features of Bi-GRU

The sort of Feature Bi-GRU makes use of contextual and sequential representations that are retrieved from Bi-GRU layers. A Bi-GRU is trained by subjecting it to vast amount of text, which enables it to understand the relationships between words and how the order in which they appear in a sequence might alter their meaning. Bi-GRUs consider data from both the start and end of a sentence, offering a more thorough comprehension of context [18].

3.3.3. CNNs for Text Analysis and Feature Engineering

The Type of Feature N-grams and local contextual representations obtained through 1D convolutional filtering are used by CNNs. Text is displayed as a series of numerals in CNN. It works with this sequence by using 1D convolutional filters that scan it in search of word or n-gram patterns [37].

3.4 Breakdown of the Prototype Design

3.4.1 TB-PLMs DistilBERT, ALBERT, and RoBERT a

DistilBERT, a transformer-based architecture, was created in 2019. It is significantly lighter, faster, and smaller than the traditional BERT variant [38]. In addition to eliminating word fragment embedding and difficulties with fixed input length size, this technique has low computing and resource usage [39]. In 2020, Google launched ALBERT, a model that builds upon the pre-trained BERT model. This model, which arrived in 2018, is a more lightweight version of the BERT. ALBERT pre-trains on text data using the transformer architecture. Among the NLP designs, Facebook AI Research (FAIR) created RoBERTa in 2019 [40]. Building on BERT's language masking technique, a Robustly Enhanced BERT Pretraining Method (RoBERTa) alters important BERT parameters. Despite BERT's 16 GB corpus, it was pretrained on 160 GB corpora over a longer period. In comparison to BERT, this enables RoBERTa representations to be applied to downstream tasks significantly more effectively [5].

3.4.2 Hybrid TB-PLMs with Bi-GRU and CNN BiGRU

The sequence of data utilizing gates to control information flow throughout the network is processed by Bi-GRU. Bi-GRU uses two gates as Update gate, reset gate and while processing a sequence in forward GRU and backward GRU units, these gates decide which new data to incorporate and what necessary data from previous periods (memory). The results from both are then combined at every stage of Bi-GRU which is as mentioned in the equations from (1)-(4). This gives information on the relationship between words in the past and the future, helps to understand the meaning and interdependence of individual words [41]. The model architecture of DistilBERT, ALBERT and RoBERTa with BiGRU are as shown in Figure. 5.

- Update Gate: Regulates the amount of fresh data that is added and the amount of historical data that is retained.

$$z_t = \sigma(W_z.\left[h_{t_{-1}}, x_t\right] + b_r) \quad (1)$$

Were,

z_t = Update gate, σ= sigmoid function,

W_z= update gate weights, h_{t-1}= past memory, x_t= current input at time step t.

- Reset Gate: decides the amount of prior knowledge to be forgotten.

$$r_t = \sigma(W_r.\left[h_{t_{-1}}, x_t\right] + b_r) \quad (2)$$

were,

r_t=reset gate, W_r=reset gate weights.

- New memory content: Calculation of new memory content

$$\widetilde{h_t} = (W.\left[r_t \odot h_{t_{-1}}, x_t\right] + b) \quad (3)$$

were,

$\widetilde{h_t}$ = New hidden state candidate, r_t=reset gate, h_{t-1}= previous hidden state, Θ=elementwise multiplication, x_t = present input, *W and b* = weights and bias.

- Final memory Update (h_t):

$$h_t = z_t \odot h_{t_{-1}} + (1 - z_t) \odot \widetilde{h_t} \quad (4)$$

were,

h_t = final updated hidden state at time t,

z_t = update gate, h_{t-1} = previous hidden state.

Now combining RoBERTa (TB-PLMs) with BiGRU for misinformation detection based on the given equations below from (5) - (8)

- RoBERTa embeddings: The word embeddings are obtained by passing the text data through RoBERTa representing the context-aware embeddings.

$$X_{RoBERTa} RoBERTa(text) \quad (5)$$

- BiGRU processing: Embeddings are then passed into BiGRU to process the sequence in both forward and backward direction.

$$BiGRU = \begin{cases} h_t^{forward} = GRU^{forward}(X_{RoBERTa}) \\ h_t^{backward} = GRU^{backward}(X_{RoBERTa}) \end{cases} \quad (6)$$

- Concatenating outputs: Combining forward and backward hidden states to capture the full context.

$$h_t^{final} = \left[h_t^{forward}; h_t^{backward}\right] \quad (7)$$

- Prediction of misinformation: Representation is passed through a dense layer to make whether the news is deceptive or genuine

$$y^{\wedge} = softmax(W.h_t^{final} + b) \quad (8)$$

CNN: In addition to DistilBERT, ALBERT, RoBERTa with BiGRU, the model architecture of ALBERT with CNN is also proposed and is depicted as shown in Fig. 6. Convolutional Neural Networks (CNNs) are suited for a variety of NLP tasks because they are efficient at capturing patterns and local relationships in sequences. Convolutional, pooling, and fully-connected layers are among the layers that make up. Its designs are particularly good at handling grid-like data, such as text [42]. The suggested model, which employs the CombinedCorpus dataset, is depicted in Fig. 7. It preprocesses the information and applies transformer-based DistilBERT, ALBERT, and RoBERTa models with improved feature learning i.e., BiGRU and ALBERT with CNN. After that, the train, test, and validate data sets are split.

The modelling approach involves working with DistilBERT, DistilBERT+BiGRU, ALBERT, ALBERT+CNN, ALBERT+BiGRU and RoBERTa, RoBERTa+BiGRU models as follows.

In model architectures that just employ the DistilBERT, ALBERT, and RoBERTa models, the classification output is supplied with the output layer (logits layer). The dense layer receives the classification result and uses the SoftMax function to categorize it into several classes. Two separate input layers, the input id layer and the attention mask layer handle the encoded text sequences along with the attention masks. These layers are integrated internally and supplied as inputs to the pre-trained model. Features retrieved from the text sequences are represented in the output of the transformer-based models that have already been trained. A stacked Bi-GRU network that aids in data synthesis and is utilized for further processing receives the extracted features.

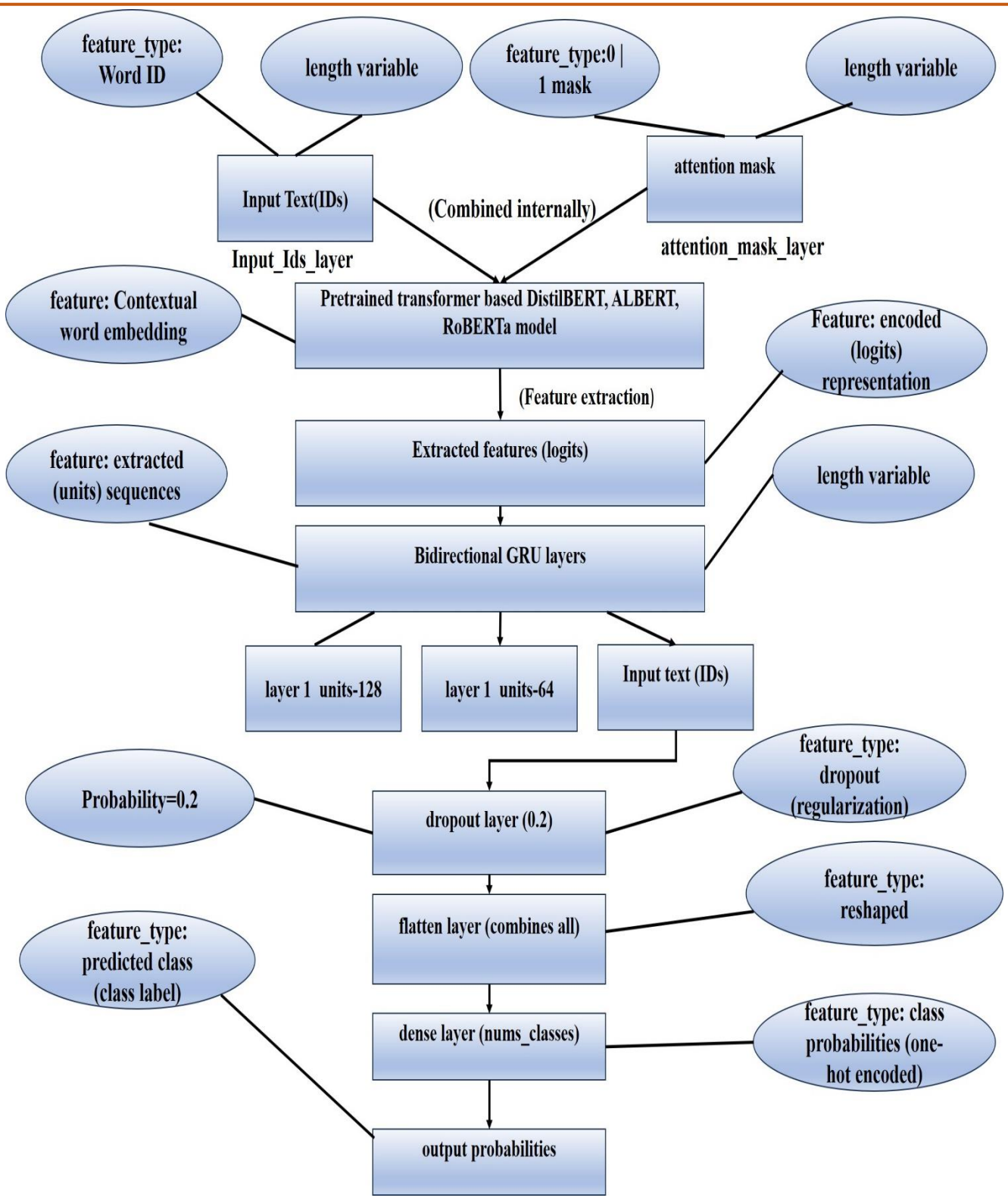

Figure 5. Model architecture of DistilBERT, ALBERT and RoBERTa with BiGRU

It makes use of dropout layer, flatten layer, and numerous GRU layers with varying numbers of units (neurons). Regularization employs the dropout layer, which randomly removes neurons during training. The output of the GRU network is reshaped by integrating all features using the flatten layer.

Additional processing is carried out using CNN on features that are retrieved from only ALBERT model. It makes use of two 1D global average pooling layers and two 1D convolutional layers. It introduces non-linearity by using filter size and extracting features using a ReLU activation function. For every feature dimension, the Global Average Pooling 1D layer determines the average value over the whole feature map.

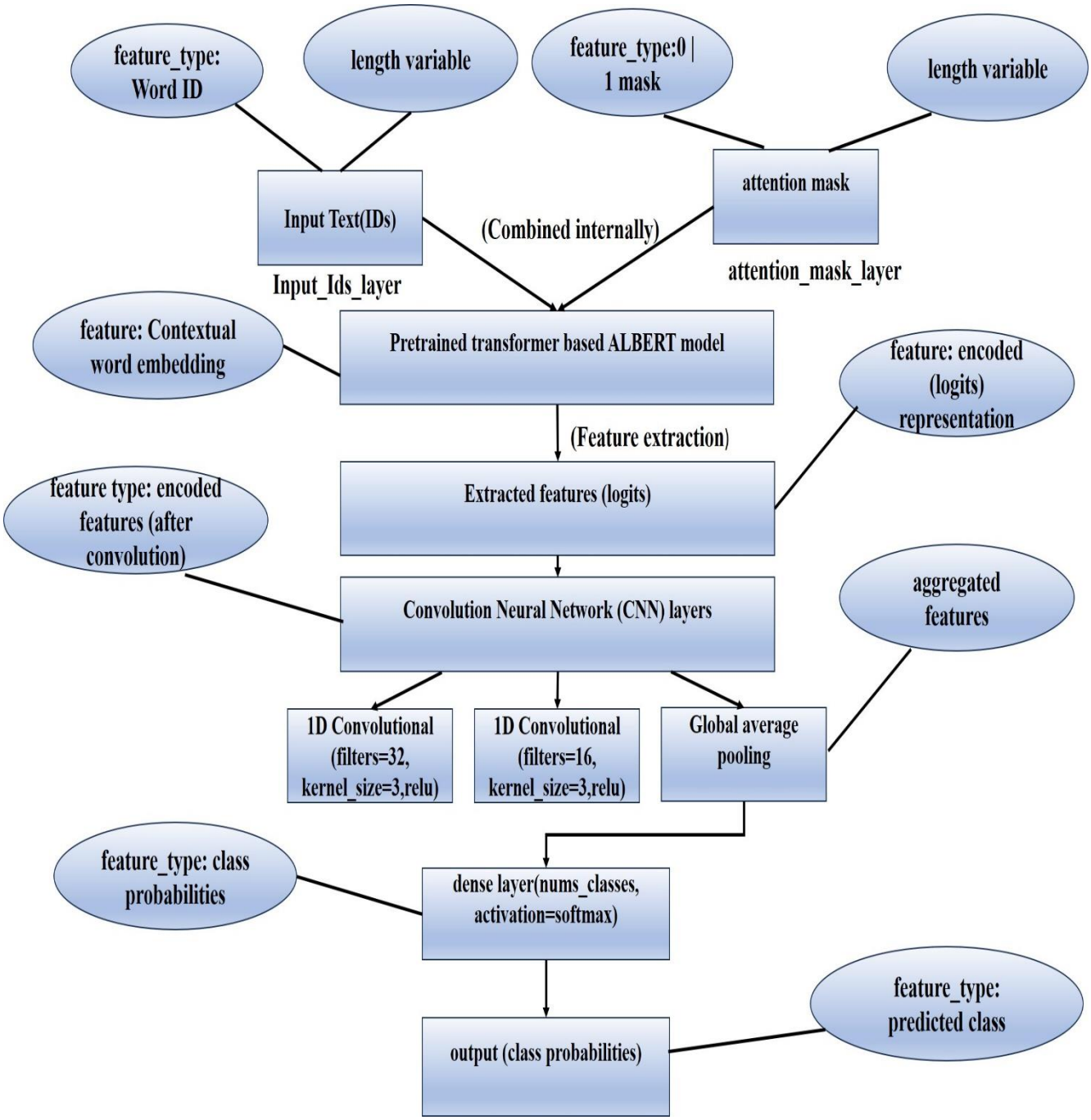

Figure 6. Model architecture of ALBERT with CNN

Table 3 below shows how proposed model architecture is different from recent hybrid models combining BERT variants with RNNs to show the novelty of proposed work through structural differences and LIME explainability.

The probability distribution across the various classes is specified in the final layer. The model is developed with an optional custom accuracy measure, a custom loss function, and built-in recall, precision, and F1-score metrics.

The workflow of the proposed models can be seen in Algorithm 1 below

Algorithm 1. TB-PLMs based Deep Neural Network models on CC dataset

Input: CombinedCorpus (CC) dataset having text and label fields: CC_D: Training sample $T_{train,}$ Validation sample $V_{valid,}$ and Testing sample $T_{test.}$

Output: Truthfulness class C= {Genuine, Deceptive} where, Probability P: C -> [0,1]

1: Read the input text from the specified location.

3: for C € preprocessed CC_D do

Apply text pre-processing and filtration. Load TB-PLMs Tokenizer objects, format the input text using

input_id and attention mask, obtain the contextual embeddings.

end for

4. for C € T_{train} do

Add Deep Neural Network structures such as BiGRU layers on DistilBERT, ALBERT, RoBERTa and CNN layers on ALBERT

Add dropout rate = 0.2, flatten and dense layer

Use Sigmoid to calculate the output probabilities of labels

end for

5. for C € T_{train}, V_{valid} do

Store loss/accuracy values, obtain training/validation loss/accuracy curves

end for

6. for C ∈ T_{test} do

Use trained model to predict the label as genuine or deceptive

end for

7. Apply LIME_explainer on the features from trained model (RoBERTa+BiGRU, DistilBERT+BiGRU) to explain the importance of each token.

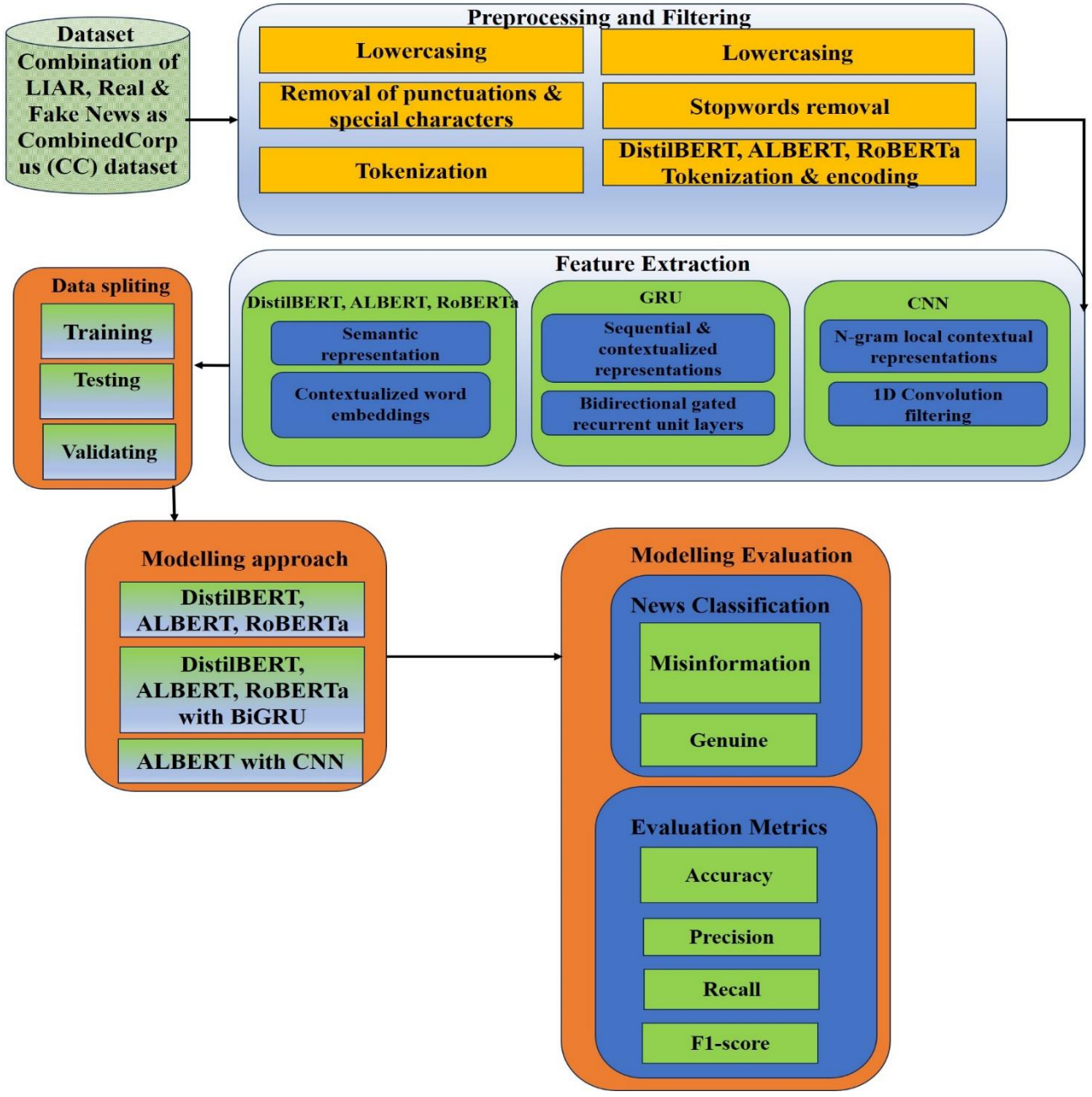

Figure 7. Architecture of the proposed model

Table 3. Key differenecs between the proposed model and recent hybrid BERT+RNN architectures

Reference	Transformer Variant	RNN (or sequential	Other layers	Difference
[43]	BERT	BiLSTM	CNN on top	Uses BiLSTM+CNN on BERT. Our model employs BiGRU with ALBERT+CNN, several transformer versions (DistilBERT, ALBERT, RoBERTa), and LIME for interpretability.
[18]	BERT, COVID-Twitter-BERT (CT-BERT)	BiGRU	----------	Uses BiGRU on top of CT-BERT or BERT. Proposed work includes hybrid models and LIME explainability, which they do not include.
[44]	RoBERTa	----------	Graph Neural Network (GNN) + Heterogeneous Attention Network (HAN)	Fuse RoBERTa with GNN+HAN (hierarchical) which is quite different from RNN. Proposed work combines RoBERTa+BiGRU (RNN), ALBERT+CNN, and LIME explainability.
[45]	Transformer (optimized)	BiGRU	Bayesian Optimization	This work uses BiGRU+transformer+Bayesian optimization. But no CNN+attention layer, and no explainability via LIME. Also, they do not compare multiple standard transformer variants like DistilBERT, ALBERT, or RoBERTa.
[46]	DistilBERT	BiGRU	Convolution+CapsNet+Selfattention	Very similar in using DistilBERT+BiGRU+ convolution+self-attention. But they do not use ALERT+CNN as an attention mechanism, nor provide LIME-based explainability, nor compare with RoBERTa / ALBERT.
Proposed work	DistilBERT; ALBERT; RoBERTa	BiGRU (for DistilBERT, ALBERT, RoBERTa), CNN (ALBERT)	ALERT (attention) + CNN; Explainability via LIME (on DistilBERT+BiGRU & RoBERTa+BiGRU); Feature importance (on DistilBERT+BiGRU)	Combines multiple transformer variants+ BiGRU, adds a hybrid CNN-attention layer, and incorporates interpretability via LIME.

4. Results and Discussions

4.1 Experimental Setup

A laptop with a Ryzen 7 5800h processor, 16GB RAM, an NVIDIA RTX 3060 GPU (Graphics Processing Units) with 6GB RAM, Python 3.9.10, and the installation of the necessary software, including CUDA 11.8, CUDNN 8.9.6, NVIDIA driver 551.86, Tensorflow 2.10, Jupyter Notebook, are used for the computations in the proposed study. A train dataset and a test dataset have been separated into an 80:20 ratio. Furthermore, a train dataset and a validation dataset were created from the training data using a validation split ratio of 0.2, and the model was run for 10 epochs with a batch size of 4. The suggested approach further makes use of a dropout

layer for regularization in the GRU network and early halting to prevent overfitting in the models. Using features from the logits layer, the study suggests an extra sub model to be added to the DistilBERT, ALBERT, and RoBERTa models. The format for the further DL model is simple as it has integrated three GRU layers for the BiGRU model and two 1D convolutional layers for the CNN model.

To prevent overfitting, the methodology incorporates a customized accuracy function and loss that utilize the L2 normalization approach for assessment purposes. By utilizing the TensorFlow framework, other metrics including precision, recall, and F1-score are also employed. The patience (number of epochs to be halted after if there is no improvement) is set to 5, and validation loss is the criterion used for early halting. An Adam optimizer with learning rates of 2e-5 (BiGRU) and 5e-5 (CNN) is used in this investigation.

4.2 Evaluation Metrics and Confusion matrix

The accuracy, precision, recall, and F1-score values of the DistilBERT, ALBERT, and RoBERTa models are compared and evaluated as per the given equations from (9)-(12)

- The percentage of correctly anticipated cases among all instances is referred to as accuracy.

$$Accuracy = TP + TN/(TP + TN + FN + FP) \quad (9)$$

- The ratio of correctly predicted instances to all expected instances for a given class is known as precision.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (10)$$

- The model's recall measures how well it detects positive examples.

$$Recall = TP/(TP + FN) \quad (11)$$

- The harmonic mean of the recall and precision scores is known as the F1-score. F1 scores are on a scale from 0 to 1, where 0 denotes a model that cannot classify any observations at all and 1 indicates a model that correctly assigns every observation to the appropriate class.

$$F1score = (2 * Precision * Recall)/(Precision + Recall) \quad (12)$$

A predictive modelling tool called a confusion matrix, is used to assess how well a categorization model performs. It gives a concise overview of the model's accuracy by showing the counts of True positive, False positive, True negative, and False negative as shown in Figure. 8, which is represented with annotated heatmaps to visually convey true/false positive and negative distribution across classes. Also, the comparison of model performance across different architectures is shown in Figure. 9, comparison of model loss is shown in Figure. 10(a) and parameters with training time in Figure. 10(b) to see the variations of different models with evaluation metrics.

Without setting the random seed, the metrics shown in the experiments initially relate to a single run for every model. Later, all models were retrained across three separate runs with distinct random seeds [1, 2, 3] to guarantee repeatability and assess statistical robustness. The averages from these runs, along with the associated standard deviations, are the performance metrics that are reported: accuracy, precision, recall, and F1-score. This proves that the performance improvements that have been seen are reliable and not the product of chance or a particular random initialization.

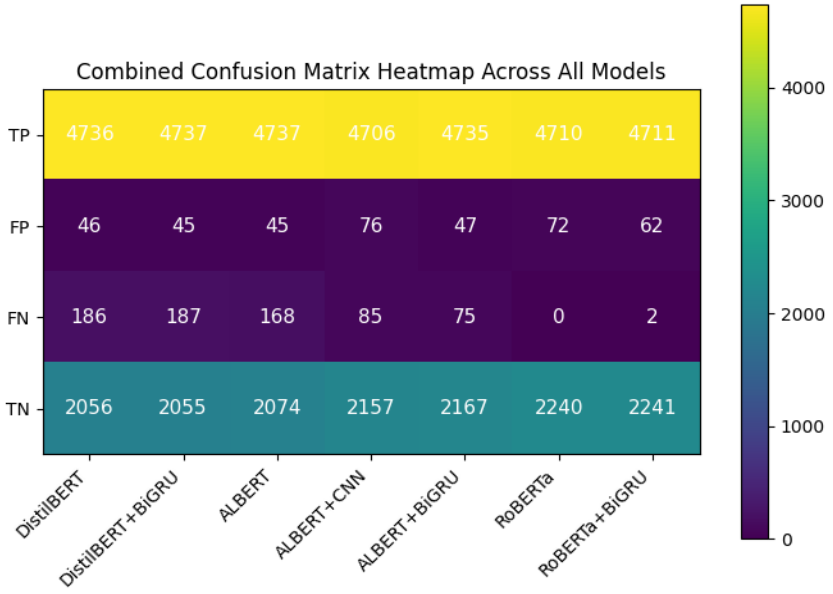

Figure 8. Confusion matrix of proposed models

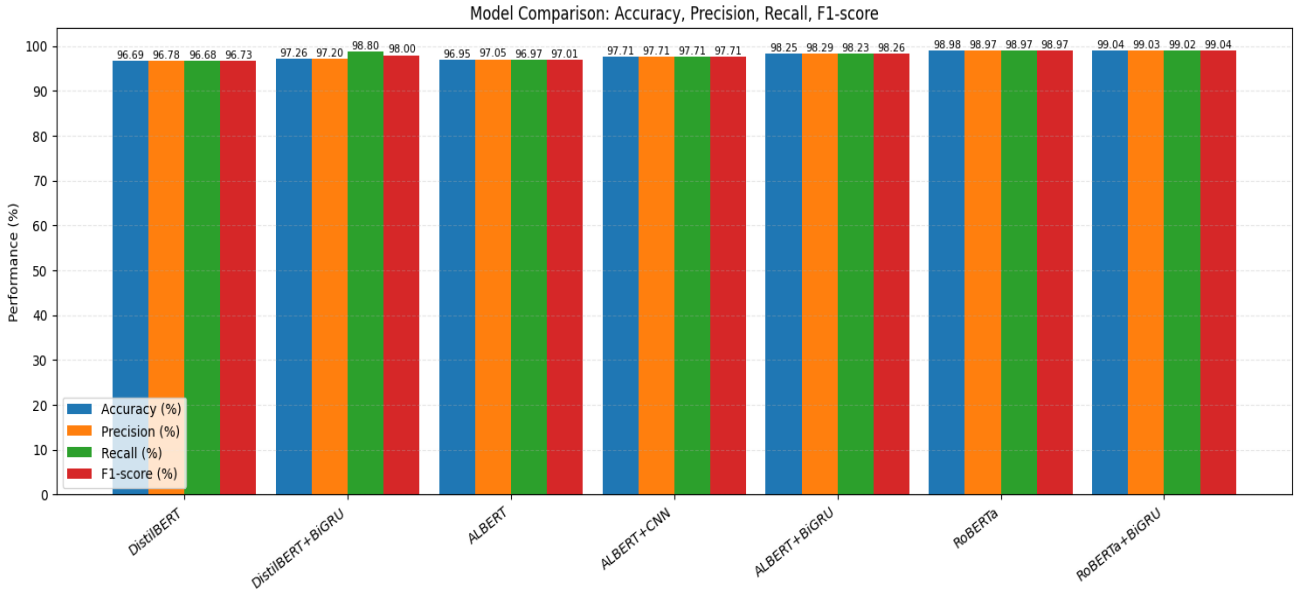

Figure 9. Comparison of model performance across different architectures

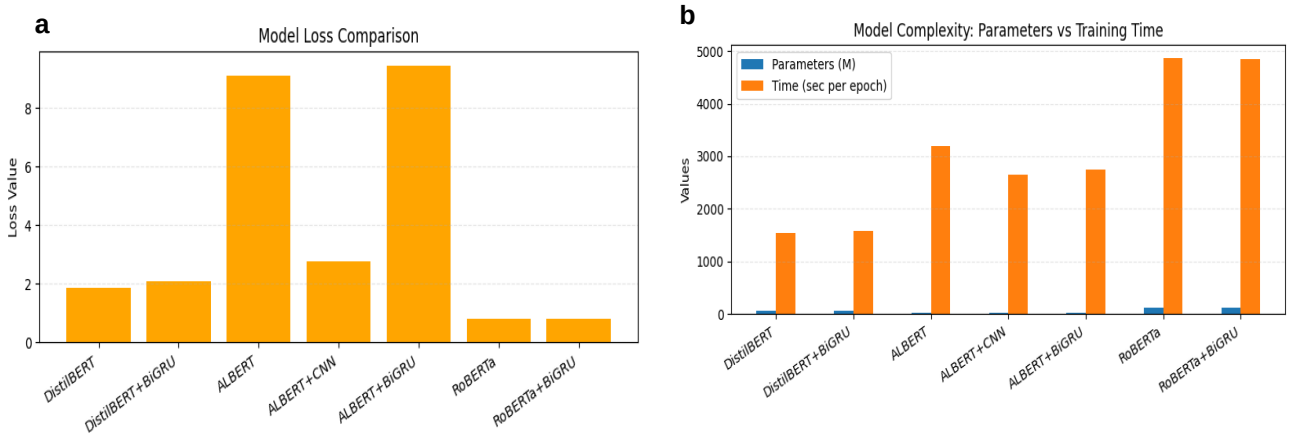

Figure 10 (a) Comparison of model loss and **(b)** parameters with training time

Table 4. Performance of various models with statistical significance measures

Model	Accuracy	Precision	Recall	F1-score	Random seed
DistilBERT	96.69±0.04	96.78±0.05	96.68±0.06	96.73±0.05	[1,2,3]
DistilBERT+ BiGRU	97.26±0.02	97.20±0.03	98.80±0.04	98.00±0.03	[1,2,3]
ALBERT	96.95±0.02	97.05±0.015	96.97±0.018	97.01±0.017	[1,2,3]
ALBERT+ CNN	97.71±0.01	97.71±0.01	97.71±0.01	97.71±0.01	[1,2,3]
ALBERT+ BiGRU	98.25±0.012	98.29±0.010	98.23±0.015	98.26±0.011	[1,2,3]
RoBERTa	98.98±0.004	98.97±0.004	98.97±0.004	98.97±0.004	[1,2,3]
RoBERTa+ BiGRU	99.04±0.012	99.03±0.014	99.02±0.015	99.04±0.013	[1,2,3]

With a mean accuracy of 99.04% on RoBERTa+BiGRU model experiment, a standard deviation of ±0.012 is obtained, the model demonstrated minor performance variability and statistically significant improvements as shown in Table 4.

4.3 Training and Validation Loss/Accuracy Curve

The term "training loss" in DL describes the error on the training dataset that shows how successfully a model including a pretrained model learns the patterns in the training data. Validation loss evaluates the model's generalizability to fresh, untested data by measuring the error on an independent validation dataset. For pretrained models to be properly fine-tuned and to avoid overfitting, a situation in which the model behaves well on training data but badly on validation data. Training accuracy, a measure of how well a pretrained model is learning from the training data. Validation accuracy illustrates how well the model generalizes on unseen data. It is possible to determine whether the pretrained model is overfitting to the training set or striking a good balance between generalization and learning by comparing these accuracies. The difference between training and validation accuracy is getting smaller, indicating that the model is learning about the underlying patterns rather than simply recalling data and improving its generalization. For the DistilBERT, ALBERT, RoBERTa, DistilBERT+BiGRU, ALBERT+BiGRU, RoBERTa+BiGRU and ALBERT+CNN models, validation accuracy shows an improvement in the effectiveness of the models on unknown data. All the suggested models' evaluation metrics are visually represented in Figure. 11, Figure. 12 and Figure. 13.

A clear illustration is provided in Table 5 about how various model configurations function under similar training conditions. It makes it possible to compare baseline transformer models with their hybrid extensions directly, demonstrating how performance can be enhanced by including sequence-modelling layers like BiGRU or CNN. This provides quantifiable proof that hybrid models outperform or match standalone transformers in fake-news detection, supporting the rationale behind our suggested architecture.

4.3.1 Ablation study

This study is made to measure each component's contribution to the Transformer-based Pretrained Language Model (TB-PLM). The findings indicate that TB-PLM alone (baseline), does not explicitly describe sequential or local patterns, although it does capture contextual embeddings. TB-PLM+BiGRU enhance the model's comprehension of context and flow by modelling long-term sequential dependencies in the text, the BiGRU layers (128 → 64 → 32, bidirectional)..

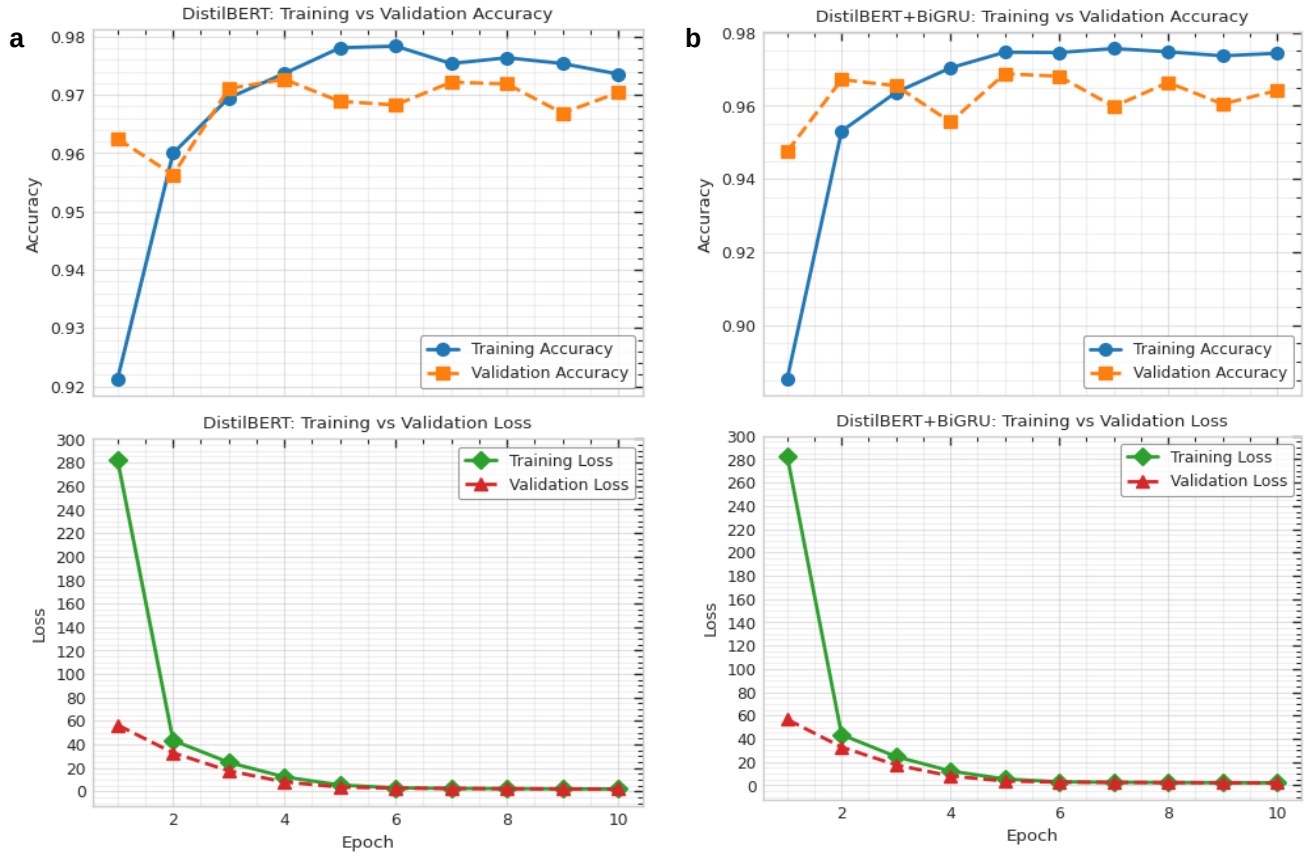

Figure 11. Training and Validation accuracy/loss curves using a) DistilBERT and b) DistilBERT+BiGRU models

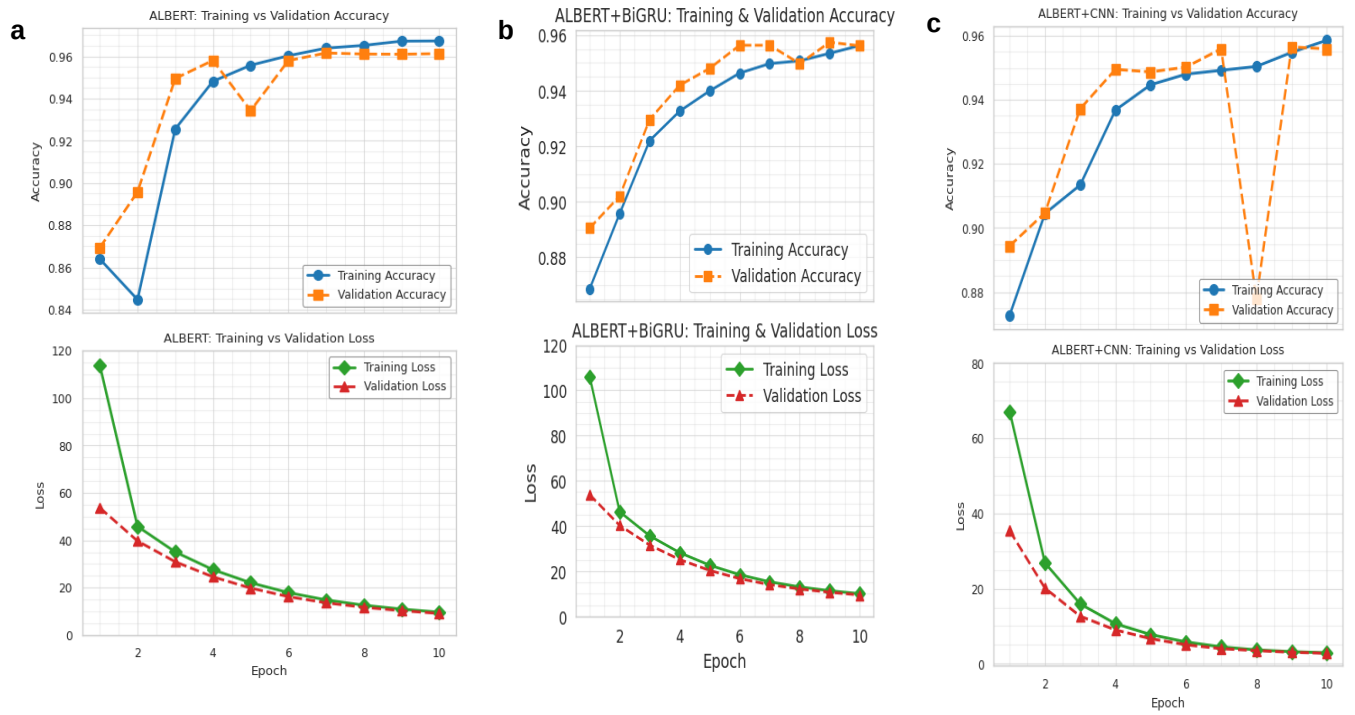

Figure 12. Training and Validation accuracy/loss curves using ALBERT, ALBERT+BiGRU and ALBERT+CNN models

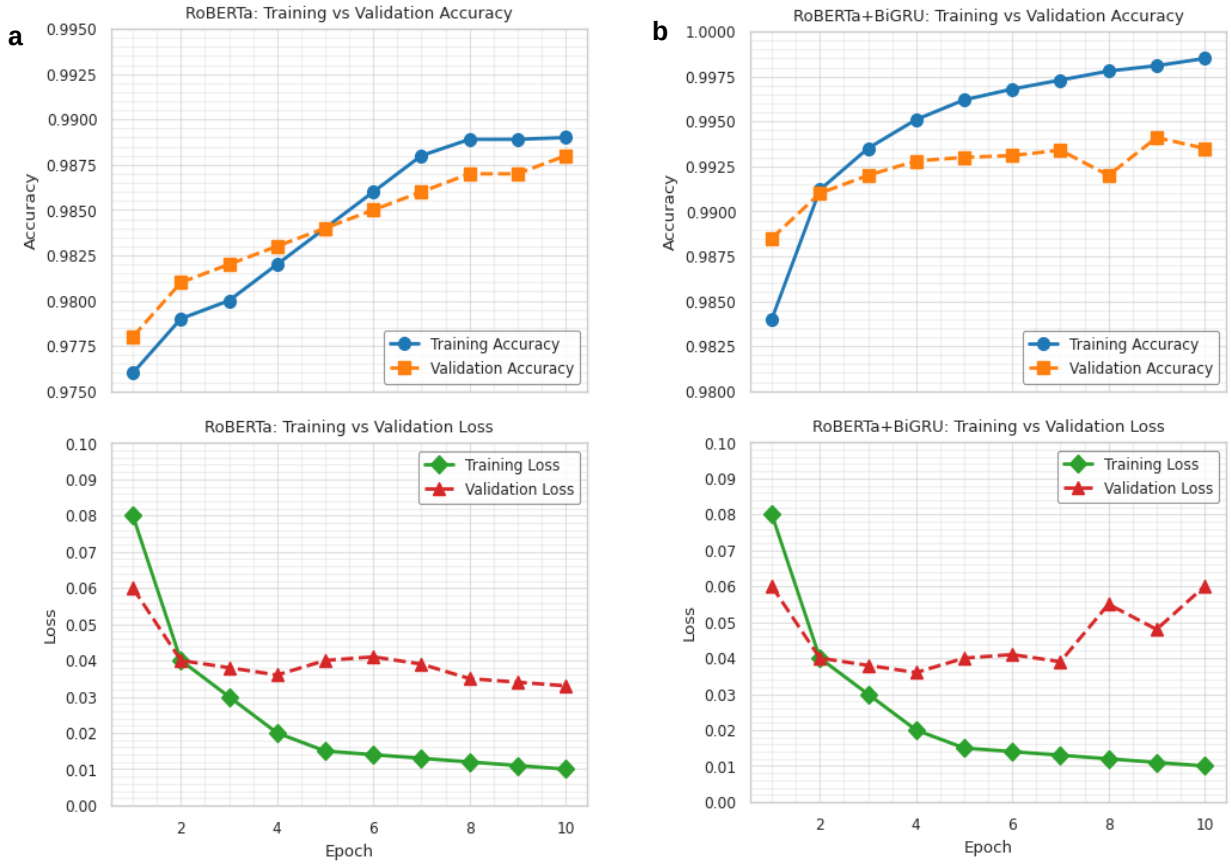

Figure 13. Training and Validation accuracy/loss curves using a) RoBERTa b) RoBERTa+BiGRU

Table 5. Impact of hyperparameters on Validation accuracy.

Common parameters used	Models	Learning rate	Layers	Rationale	Validation accuracy
Batchsize=4, Epochs=10, Regularization=0.01, ADAM optimizer, Early stopping= Patience value (5), Train, Test, validate dataset: 80%, 20% and 20% of train set respectively.	DistilBERT	2 e-5	-	Small batch for stable training and memory constraints, Sufficient for convergence based on validation performance	97.04
	DistilBERT+ BiGRU	2 e-5	128, 64	adapts learning rates for stable convergence, learning rates for transformer fine-tuning; chosen to ensure stable convergence	96.11
	ALBERT	2 e-5	-	Standard split to evaluate generalization performance.	96.13
	ALBERT+ CNN	5 e-5	Conv1D (32, k=3) →Conv1D (16, k=2) →GlobalAvgPool1D	Stops training if validation loss does not improve for 5 epochs	95.59
	ALBERT+ BiGRU	2 e-5	128,64,32	L2 regularization to prevent overfitting and improve generalization	95.64
	RoBERTa	2 e-5	-	Strong baseline transformer with improved training.	98.25
	RoBERTa+ BiGRU	2 e-5	128, 64	BiGRU enhances modeling of word order & context.	98.70

In comparison to the baseline, this results in higher recall and F1-score. TB-PLM+CNN improves the model's precision in identifying discriminative phrases by capturing local n-gram patterns, the CNN layers (Conv1D 32→16 with kernel sizes 3→2 followed by Global Average Pooling). In DistilBERT+BiGRU and RoBERTa+BiGRU models, distilBERT and RoBERTa models are contributing above 90%; in ALBERT+BiGRU and ALBERT+CNN models, ALBERT alone is contributing above 85% a validation accuracy after a single epoch providing strong semantic representations, demonstrating TBPLMs already capture highly discriminative contextual information. Subsequent epochs show gradual improvements in validation accuracy indicating BiGRU (task specific sequential enhancement) and CNN (high resolution for local pattern detection) layers contribute by refining the transformer outputs and modelling deeper feature interactions.

4.4 Explainable AI (XAI) model LIME

Users are more inclined to believe the findings and follow suggestions in deceptive news detection if they are provided with detailed reasons for why a piece of news was categorized as legitimate or deceptive. Explainability is crucial for detecting deceptive news because it fosters cooperation with human experts, increases user involvement, establishes legal compliance, boosts knowledge, and recognizes misconceptions. One method of explaining models is Local Interpretable Model-Agnostic Explanations (LIME) [47] uses models that provide probabilities as a solution for categorization issues. The importance of every

feature is explained by assigning the weights on each word showing that how important a word contributing to the performance of a model by LIME explainer as shown Figure. 14a. The bar chart in Figure. 14b shows float point numbers on the horizontal bars representing the relative importance of these features (green for class 1 and red for class 0). LIME [48] maintains the explanatory ability of significant features regardless of the chosen classifier running independently of the model used. For a given text input, LIME fits a local linear model around the prediction based on the equation (13)-(17) given below

- Local Linear Model for LIME is given as follows

$$f(y) = \sum_{i=0}^{n} \beta_i \Phi_i(y) \quad (13)$$

Were,

F(y)= prediction of a model on the input 'y'

β_i= weights assigned for each feature $\phi_i(y)$

$\phi_i(y)$= feature transformations such as words or word embeddings.

- Euclidean Distance metric d(y, y_j) = $\sqrt{\sum_{i=1}^{n}(y_i - y_{ji})^2}$ (14)

to calculate the weights w_j.

- Scaling parameter or kernel width σ in the weight function $w_j = \exp(-\frac{d(y-y_j)^2}{\sigma^2})$ is given as,

$$\sigma = \alpha.\text{mean}\,(d\,(y, yj)) \quad (15)$$

Were,

α is a tuning parameter

mean(d(y, y_j)) – median distance between the initial inputs and its perturbations

- Sampling of perturbation, P(drop word)= p_{drop}
- Simplicity to regularizations

β^=argmin$_\beta$ $\sum_{j=1}^{m} w_j(f(y_i) - \sum_{i=0}^{n} \beta_i \phi_i(y_j))$^2 + $\lambda \sum_{i=0}^{n} |\beta_i|$ (16)

Were,

λ= regularization parameter, $\sum_{i=0}^{n} |\beta_i|$ penalize the bigger values of β_i

- Selection of important features is given as $S_i = \sum_{j=1}^{m}(\frac{\partial f(y_j)}{\partial \phi_i})$^2 for feature rankings (17)

were,

S_i – importance score for the feature ϕ_i,

$\sum_{j=1}^{m}(\frac{\partial f(y_j)}{\partial \phi_i})$ – summation of m perturbed instances y_j, over the partial derivative of the model's prediction function f(y_j).

The text explainer finds the top words which primarily drive the model to make the classification decision providing intuitive model behavior. This maps with the original class label providing individual feature relevance and high feature contribution in the final prediction by highlighting the text. The word "[CLS]" is classification token serves as an aggregate representation of the entire input sequence and "[SEP]" is a separator token used to mark the end of a sentence which help the model process and classify text correctly. The words "##line" might be part of "headline," and "##ock" could be part of "stock". These words are required to show as how transformer-based models (like DistilBERT) handle words and are important for representing incomplete or rare words in the input Also they are the integral part of the model's architecture and tokenization process.

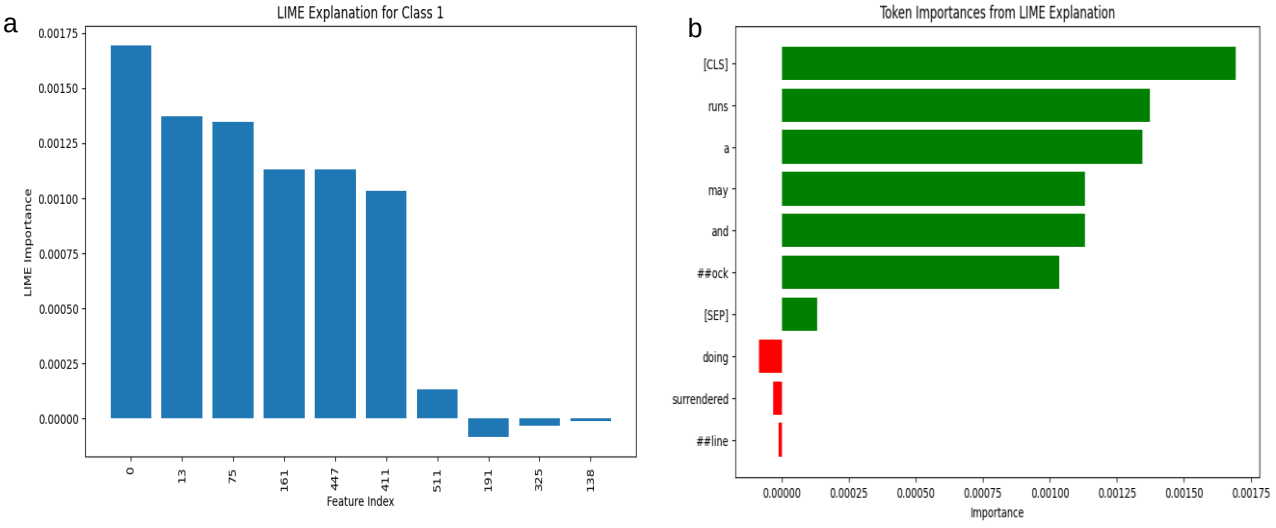

Figure 14 a) LIME Explainer for class 1 with a feature index and importance. **b)** LIME interpretations include prediction probabilities for both classes (0 and 1) according to the highlighted colour and grade given for each phrase in the sentence text on DistilBERT+BiGRU

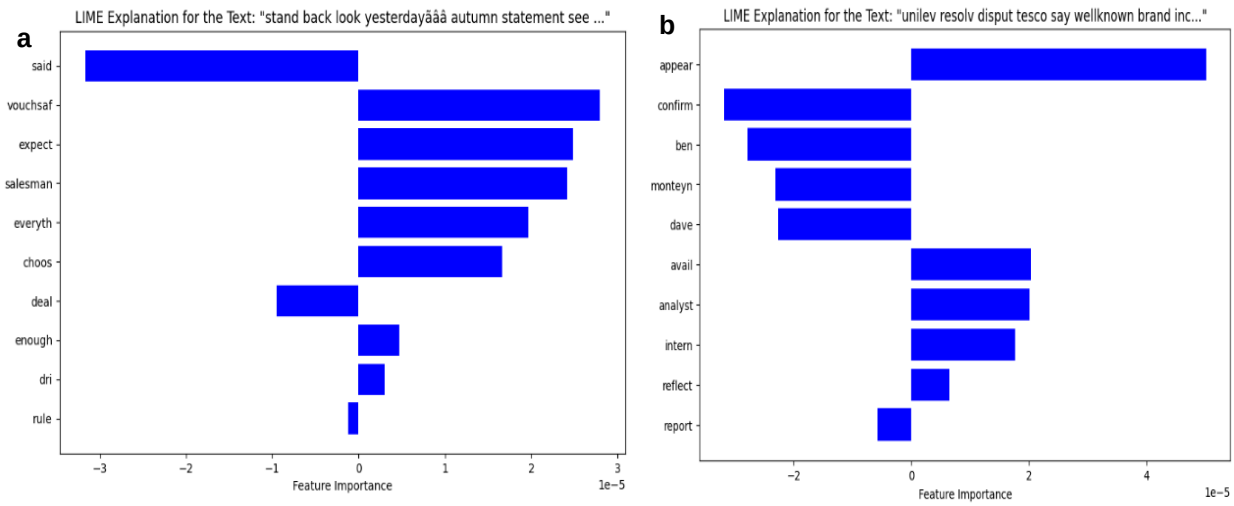

Figure 15. LIME Explainer for two sample text sentences on RoBERTa+BiGRU model. **a)** LIME Explainer sample1, **b)** LIME Explainer sample2

Rest of the word's importance is given based on the weights. The bar graph in Figure. 14b displays floating values on the horizontal lines showing the corresponding importance of features such as green for class 1 and red for class 0.

Figure. 15a and 15b provides two text samples and explains the feature importance of indicational words appearing in the text corpus on RoBERTa+BiGRU model. To increase trust for application and implementation, LIME usually serves as an explanation tool for both experts and novice users with a range of explainability criteria. To provide in-built model behavior, the text explainer identifies the top phrases that influence the model's categorization decision. By emphasizing the text's content, this correlates with the initial class label, offering unique feature significance and high feature contribution in the final prediction. For detecting deceptive information, the model can differentiate between legitimate and deceptive statements.

5. Comparative Analysis

The suggested work has attained 98.98% accuracy when used, but the current research is based on different transformer-based pretrained models for deceptive news identification, which scoreless accurately altogether. RoBERTa, the model uses a large CC dataset with 34564 train records, 8640 validate records, and 7024 test records to learn the sequential patterns. These patterns are found in the semantic representations of the data, and they can provide clues about the accuracy of news articles.

Comparatively, the suggested work scores higher accuracy than the current DistilBERT research work. A working environment for the proposed work is created by considering the dataset of previous work from [6, 49]. The NELA-GT-2019 dataset, which comprises 28743 test records and 60087 train records, is one of the datasets used in [6]. The DistilBERT model achieves an accuracy of 91.64%, while the DistilBERT+BiGRU model achieves a 96.69% accuracy. The Constraint@AAAI2021-COVID-19 dataset, which is the basis for [49], consists of 6420 train records and 2140 validated data points on English news. The DistilBERT model achieves an accuracy of 90.92%, while the DistilBERT+BiGRU model achieves an accuracy of 52.34%.

On ALBERT+BiGRU layers, the ALBERT model obtained 98.25% accuracy and 96.95% accuracy. A working environment for the proposed work is created by considering the dataset of previous work from [50, 51]. The Gossipcop dataset used in [50] has 15497 train records, 3321 validate records, and 3322 test records on English news. These records are used to run the ALBERT model, which achieves an accuracy of 81.13%, and the ALBERT+BiGRU model, which achieves an accuracy of 75.16%. A dataset utilized in [51] is derived from Constraint@AAAI2021 and the ALBERT model is run with 6420 train records, 2140 validate records, and 2140 test data to achieved an accuracy of 92.66%. And, the ALBERT+BiGRU model achieved an accuracy of 52.37%.

RoBERTa model achieves an accuracy of 98.98% accuracy whereas RoBERTa+BiGRU achieves an accuracy of 99.04%. A working environment for the proposed work is created by considering the dataset of previous work from [33] where the dataset considered is Fake or real News, LIAR and Combined Corpus with a total record of 6335, 12791 and 79548 respectively related to Politics, economy, investigation, health, sports, entertainment. The RoBERTa model achieves an accuracy of 96.00% and 97.25% on Combined corpus when their dataset is run on working environment of the proposed work. Table 6 gives the comparison of pretrained model DistilBERT, ALBERT and RoBERTa accuracy with existing work by running the datasets of existing work in working environment of proposed model.

Table 6. Comparison of pretrained model DistilBERT, ALBERT and RoBERTa accuracy with existing work by running the datasets of existing work in working environment of proposed model

Ref.	Dataset	Model	Accuracy (%)
[6]	NELA-GT-2019 dataset	DistilBERT	91.64
		DistilBERT + BiGRU	96.69
[49]	Constraint@AAAI2021-COVID-19 dataset	DistilBERT	90.92
		DistilBERT + BiGRU	52.34
		ALBERT	81.13
		ALBERT + BiGRU	75.16
[51]	Constraint@AAAI2021	ALBERT	92.66
		ALBERT + BiGRU	52.37
[19]	Fake or real News, LIAR and Combined Corpus	RoBERTa	96.00
		RoBERTa +BiGRU	97.25
Proposed Work	CombinedCorpus (CC)	Support Vector Machine (SVM)	95.82
		Logistic Regression (LR)	95.61
		Long-short Term Memory (LSTM)	95.92
		DistilBERT	96.69
		DistilBERT+ BiGRU	97.26
		ALBERT	96.95
		ALBERT+ BiGRU	98.25
		RoBERTa	**98.98**
		RoBERTa+BiGRU	**99.04**

In contrast to earlier studies, proposed work presents a systematic examination of hybrid architectures that integrate recurrent and convolutional refinement layers like BiGRU and CNN with several compact transformer families, including DistilBERT, ALBERT, and RoBERTa. To capture sequential dependencies that traditional pooling could overlook, RoBERTa with a Bidirectional GRU layer re-encodes transformer outputs, resulting in more discriminative phrase representations. Furthermore, a staged fine-tuning method is utilized wherein RoBERTa is frozen while the BiGRU and classifier are trained, followed by a progressive unfreezing with a decreased learning rate. This sets the approach apart from transformer-only or independently trained RNN and transformer models by increasing training stability and decreasing overfitting. Previous studies concentrate mostly on BERT-based models with dense layers [10]. While authors in [17] assess deep learning models (CNN, LSTM, BERT) without transformer, RNN hybridization. In [20], authors do not use efficiency-oriented model combinations or recurrent refinement, in their investigation of transformer ensembles for Hindi fake news detection. In the proposed model, stronger sequential dependency capture is made possible by combining the robust contextual encoding of RoBERTa with a bidirectional GRU layer, especially the RoBERTa+BiGRU model. This design offers a novel contribution beyond previous methods by achieving greater test accuracy while maintaining computational efficiency.

6. Conclusion and Future Scope

The research study investigates the efficacy of a transformer-based English deceptive news detection system on a political news dataset. A comprehensive experimental computation has demonstrated that the proposed model can identify and classify deceptive information with an NVIDIA RTX 3060 GPU. Using transformer designs, the DistilBERT, ALBERT, and RoBERTa models successfully capture semantic linkages and contextual relationships to allow accurate predictions, which are then fed as an input to CNN and BiGRU deep learning models. There includes a thorough discussion of the importance of feature engineering, including text preparation techniques. Experimental findings show that the proposed approach exceeds previous efforts in terms of accuracy, precision, recall, and F1-score when DistilBERT, ALBERT, and RoBERTa models are employed in conjunction with BiGRU, and worked on ALBERT+CNN models. Text and other sequential information can be processed efficiently via the BiGRU paradigm. Using forward as well as backward GRU units, which capture the past and future dependencies of a word in the sequence, it checks long-term dependencies inside a sequence. CNN is better at classifying data by identifying regional trends. In comparison to all other models, the RoBERTa+BiGRU model outperforms them all in identifying deceptive information in news reports, with an accuracy percentage of 99.04%. Also, an XAI LIME is used with DistilBERT+BiGRU and RoBERTa+BiGRU to explain about influence of words used in the text based on labelled features.

Furthermore, deep learning models on deceptive information detection of English news on large corpus can be used to create more sophisticated transformer-based pretrained language models with more parameters, such as ELECTRA, ELMo, BERT, and XlNet embeddings. Also, proposed work will be extended to work with other XAI like SHAP on various pretrained transformer-based models.

References

[1] N.F. Baarir, A. Djeffal, (2021) Fake news detection using machine learning. In 2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH), IEEE, Algeria. https://doi.org/10.1109/IHSH51661.2021.9378748

[2] R.M. Johnson, Social Media and Free Speech: A Collision Course That Threatens Democracy. Ohio Northern University Law Review, 49(2), (2023) 461- 487.

[3] S.Z. Akbar, A. Panda, J. Pal, (2024) Political hazard: Misinformation in the 2019 Indian general election campaign. In Political Campaigning in Digital India, Routledge.

[4] W. Y. Wang, LIAR, LIAR pants on fire. A new benchmark dataset for fake news detection. arXiv preprint arXiv, 1705.00648. https://doi.org/10.48550/arXiv.1705.00648

[5] A. Hande, K. Puranik, R. Priyadharshini, S. Thavareesan, B.R. Chakravarthi, (2021). Evaluating pretrained transformer-based models for COVID-19 fake news detection. In 2021 5th international conference on computing methodologies and communication (ICCMC), IEEE, India. https://doi.org/10.1109/ICCMC51019.2021.9418446

[6] S. Raza, C. Ding, Fake news detection based on news content and social contexts: a transformer-based approach. International Journal of Data Science and Analytics, 13(4), (2022) 335-362. https://doi.org/10.1007/s41060-021-00302-z

[7] R. Mohawesh, S. Maqsood, Q. Althebyan, Multilingual deep learning framework for fake news detection using capsule neural network. Journal of Intelligent Information Systems, 60(3), (2023) 655-671. https://doi.org/10.1007/s10844-023-00788-y

[8] S. Kula, M. Choras, R. Kozik, (2021). Application of the BERT-based architecture in fake news detection. In 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020), CISIS 2019. Advances in Intelligent Systems and Computing, Springer, Cham. https://doi.org/10.1007/978-3-030-57805-3_23

[9] H. Jwa, D. Oh, K. Park, J.M. Kang, H. Lim, exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). Applied Sciences, 9(19), (2019) 4062. https://doi.org/10.3390/app9194062

[10] F. Al-Quayed, D. Javed, N.Z. Jhanjhi, M. Humayun, T.S. Alnusairi, Optimizing Fake News Detection. A Hybrid Transformer-Based Model for Enhanced Performance, IEEE Access, 12, (2024) 160822 – 160834. https://doi.org/10.1109/ACCESS.2024.3476432

[11] P. Gupta, S. Gandhi, B.R. Chakravarthi, Leveraging transfer learning techniques-bert, roberta, albert and distilbert for fake review detection. In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, (2021) 75-82. https://doi.org/10.1145/3503162.3503169

[12] S.R. Sahoo, B.B. Gupta, Multiple features-based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing, 100, (2021) 106983. https://doi.org/10.1016/j.asoc.2020.106983

[13] S. Hakak, M. Alazab, S. Khan, T.R. Gadekallu, P.K.R. Maddikunta, W.Z. Khan, An ensemble machine learning approach through effective feature extraction to classify fake news. Future Generation Computer Systems, 117, (2021) 47-58. https://doi.org/10.1016/j.future.2020.11.022

[14] A. Jarrahi, L. Safari, (2023). Evaluating the effectiveness of publishers' features in fake news detection on social media. Multimedia Tools and Applications, 82(2), (2023) 2913-2939. https://doi.org/10.1007/s11042-022-12668-8

[15] T. Pavlov, G. Mirceva, (2022), May Covid-19 fake news detection by using bert and roberta models. 45th Jubilee International Convention on Information. Communication and Electronic Technology (MIPRO), IEEE, Croatia. https://doi.org/10.23919/MIPRO55190.2022.9803414

[16] C. Busioc, V. Dumitru, S. Ruseti, S. Terian-Dan, M. Dascalu, T. Rebedea, (2022). What are the latest fake news in romanian politics? an automated analysis based on bert language models. In Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education. Proceedings of the 6th International Conference on Smart Learning Ecosystems and Regional Development, Springer Singapore. https://doi.org/10.1007/978-981-16-3930-2_16

[17] A. Wani, I. Joshi, S. Khandve, V. Wagh, R. Joshi, (2021). Evaluating deep learning approaches for covid19 fake news detection. In International Workshop on Combating On line Ho stile Posts in Regional Languages during Emergency Situition Springer International Publishing, Springer,

Cham. https://doi.org/10.1007/978-3-030-73696-5_15

[18] J. Alghamdi, Y. Lin, S. Luo, (2023). Towards COVID-19 fake news detection using transformer-based models. Knowledge-Based Systems, 274, (2023) 110642. https://doi.org/10.1016/j.knosys.2023.110642

[19] J.Y. Khan, Md.T.I. Khondaker, S. Afroz, G. Uddin, A. Iqbal, A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, (2021) 100032. https://doi.org/10.1016/j.mlwa.2021.100032

[20] M. Aman, Large language model based fake news detection. Procedia Computer Science, 231, (2024) 740-745. https://doi.org/10.1016/j.procs.2023.12.144

[21] B.Chen, B. Chen, D. Gao, Q. Chen, C. Huo, X. Meng, W. Ren, Y. Zhou,(2021).Transformer-based language model fine-tuning methods for COVID-19 fake news detection. In International Workshop onCombating on line Ho st ile Posts in Regional Languages during Emergency Situition. Springer, Cham. https://doi.org/10.1007/978-3-030-73696-5_9

[22] S. Kula, R. Kozik, M. Choras, M. Woźniak, (2021). Transformer based models in fake news detection. Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, Springer, Cham.

[23] M. Samadi, M. Mousavian, S. Momtazi, Deep contextualized text representation and learning for fake news detection. Information processing & management, 58(6), (2021) 102723. https://doi.org/10.1016/j.ipm.2021.102723

[24] O. Bashaddadh, N. Omar, M. Mohd, M.N. Akmal Khalid, Machine Learning and Deep Learning Approaches for Fake News Detection. A Systematic Review of Techniques, Challenges, and Advancements, IEEE Access, 13, (2025) 90433 – 90466. https://doi.org/10.1109/ACCESS.2025.3572051

[25] Bo Hu, Z. Mao, Y. Zhang, An overview of fake news detection; from a new perspective. Fundamental Research, 5(1), (2025) 332-346. https://doi.org/10.1016/j.fmre.2024.01.017

[26] K. Irfan, M. Wasim, S. Safdar, A. Rehman, M.U. Ghani, (2025). XFND: Explainable Fake News Detection using a Hybrid DistillBERT and BiLSTM. In 2025 International Conference on Emerging Technologies in Electronics, Computing, and Communication (ICETECC), IEEE, Pakistan. https://doi.org/10.1109/ICETECC65365.2025.11070272

[27] M. Al-alshaqi, D.B. Rawat, C. Liu, A BERT-Based Multimodal Framework for Enhanced Fake News Detection Using Text and Image Data Fusion. Computers, 14(6), (2025) 237. https://doi.org/10.3390/computers14060237

[28] A. De, D. Bandyopadhyay, B. Gain, A. Ekbal, A transformer-based approach to multilingual fake news detection in low-resource languages. Transactions on Asian and Low-Resource Language Information Processing, 21(1), (2021) 1-20. https://doi.org/10.1145/3472619

[29] E. Essa, K. Omar, A, Alqahtani, Fake news detection based on a hybrid BERT and LightGBM models. Complex & Intelligent Systems, 9(6), (2023) 6581-6592. https://doi.org/10.1007/s40747-023-01098-0

[30] S. Kula, R. Kozik, M. Choras, Implementation of the BERT-derived architectures to tackle disinformation challenges. Neural Computing and Applications, 34(23), (2022) 20449-20461. https://doi.org/10.1007/s00521-021-06276-0

[31] S. Nwaiwu, N. Jongsawat, A. Tungkasthan, A. Decoding Disinformation: A Feature-Driven Explainable AI Approach to Multi-Domain Fake News Detection. Applied Sciences, 15(17), (2025) 9498. https://doi.org/10.3390/app15179498

[32] Saadi, Abdelhalim, H. Belhadef, A. Guessas, O. Hafirassou, Enhancing Fake News Detection with Transformer Models and Summarization. Engineering, Technology & Applied Science Research, 15(3), (2025) 23253-23259. https://doi.org/10.48084/etasr.10678

[33] A. Praseed, J. Rodrigues, P.S. Thilagam, Hindi fake news detection using transformer ensembles. Engineering Applications of Artificial Intelligence, 119, (2023) 105731. https://doi.org/10.1016/j.engappai.2022.105731

[34] H.R. LekshmiAmmal, A.K. Madasamy, A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. Journal of Big Data, 12(1), (2025) 46. https://doi.org/10.1186/s40537-025-01093-x

[35] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash, S. Jadhav, Explainable Multilingual and Multimodal Fake News Detection: Towards Robust and Trustworthy AI for Combating Misinformation. Frontiers in Artificial Intelligence, 8, (2025) 1690616.

[36] X. Men V.Y. Mariano, Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19. International Journal of Modern Education and Computer Science (IJMECS), 16(1), (2024) 11-22.

[37] M. Samadi, S. Momtazi, Fake news detection: deep semantic representation with enhanced feature engineering. International Journal of Data Science and Analytics, 20(2), (2025) 325-336. https://doi.org/10.1007/s41060-023-00387-8

[38] V. Sanh, L. Debut, J. Chaumond, T. Wolf, (2019) DistilBERT, a distilled version of BERT: smaller,

faster, cheaper and lighter. arXiv preprint arXiv 1910.01108. https://doi.org/10.48550/arXiv.1910.01108

[39] S. Silalahi, T. Ahmad, H. Studiawan, (2022). Named entity recognition for drone forensic using Bert and DistilBERT. In 2022 International Conference on Data Science and Its Applications (ICoDSA), IEEE, Indonesia. https://doi.org/10.1109/ICoDSA55874.2022.9862916

[40] S.F.N Azizah, H.D. Cahyono, S.W. Sihwi, W. Widiarto, (2023). Performance analysis of transformer based models (BERT, ALBERT, and RoBERTa) in fake news detection. In 2023 6th International Conference on Information and Communications Technology (ICOIACT), IEEE, Indonesia. https://doi.org/10.1109/ICOIACT59844.2023.10455849

[41] H. ELFAIK, Automatic detection of fake news using gated recurrent unit deep model. Procedia Computer Science, 233, (2024) 474-480. https://doi.org/10.1016/j.procs.2024.03.237

[42] H. Saleh, A. Alharbi, S.H. Alsamhi, OPCNN-FAKE: Optimized convolutional neural network for fake news detection. IEEE Access, 9, (2021) 129471-129489. https://doi.org/10.1109/ACCESS.2021.3112806

[43] Y. Wang, Y. Zhang, X. Li, X. Yu, (2021) Covid-19 fake news detection using bidirectional encoder representations from transformers based models. arXiv preprint arXiv:2109.14816. https://doi.org/10.48550/arXiv.2109.14816

[44] T. Shwetha, R. Buvanaa, J. Jayabharathy, I. Sivasakthi, R.H. Sai, Fake News Detection Using Hybrid Transformer-Based Model. IJSAT-International Journal on Science and Technology, 16(2), (2025) https://doi.org/10.71097/IJSAT.v16.i2.5305

[45] Leveraging Bayesian optimization and bidirectional recurrent unit. arXiv preprint arXiv:2502.09097. https://doi.org/10.48550/arXiv.2502.09097

[46] M.I. Nadeem, S.A.H. Mohsan, K. Ahmed, D. Li, Z. Zheng, M. Shafiq, S.M. Mostafa, HyproBert: A fake news detection model based on deep hypercontext. Symmetry, 15(2), (2023) 296. https://doi.org/10.3390/sym15020296

[47] G. Joshi, A. Srivastava, B. Yagnik, M. Hasan, Z. Saiyed, L.A. Gabralla, K. Kotecha, Explainable misinformation detection across multiple social media platforms. IEEE Access, 11, (2023) 23634-23646. https://doi.org/10.1109/ACCESS.2023.3251892

[48] V. Dua, A. Rajpal, S. Rajpal, M. Agarwal, N. Kumar, I-flash: Interpretable fake news detector using lime and shap. Wireless Personal Communications, 131(4), (2023) 2841-2874. https://doi.org/10.1007/s11277-023-10582-2

[49] A.U. Hussna, I.I. Trisha, M.S. Karim, M.G.R. Alam, (2021) COVID-19 fake news prediction on social media data. IEEE Region 10 Symposium (TENSYMP), IEEE, Korea. https://doi.org/10.1109/TENSYMP52854.2021.9550957

[50] K. Pelrine, J. Danovitch, R. Rabbany, The surprising performance of simple baselines for misinformation detection. In Proceedings of the web conference, (2021) 3432-3441. https://doi.org/10.1145/3442381.3450111

[51] S. Malla, P.J.A Alphonse, Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news. The European Physical Journal Special Topics, 231(18), (2022) 3347-3356. https://doi.org/10.1140/epjs/s11734-022-00436-6

Acknowledgement

We would like to acknowledge University Visvesvaraya College of Engineering, Bangalore University for providing platform for conducting the research experiments.

Authors Contribution Statement

Arati M Chabukswar: Conceptualization, Methodology, Investigation, Writing-Original Draft, Visualization, Project administration. P Deepa Shenoy: Validation, Writing-Review & Editing, Supervision. S.M. Dasharath: Writing-Review & Editing, Supervision, Visualization. K.R. Venugopal: Writing-Review & Editing, Supervision. All the Authors Read and Approved the Final Version of the Manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

On the online community repository GitHub (https://github.com/JunaedYounusKhan51/FakeNewsDetection), this dataset can be accessed by the public under the FakeNewsDetection dataset.

Has this article screened for similarity?

Yes

About the License