Fake News Detection Using Bert-Based Language Model- With Explainable AI

Sandeep Yadav[1], Saud Haseab[2], Shiv Shankar Yadav[3], Mrs. Vaishali Singh[4]
[1,2,3]*School of Computer Applications, Babu Banarasi Das University, Lucknow, Uttar Pradesh, India*
[4]*Assistant Professor, School of Computer Applications, Babu Banarasi Das University, Lucknow, Uttar Pradesh, India*
doi.org/10.64643/IJIRT209638

***Abstract*—social media has become one of the fastest means of spreading false information, thus there is an urgent need to develop automatic fact-checking mechanisms. The prior fake news detection systems were largely based upon manually engineered lexical features and classical machine learning classification techniques. Recent studies in deep learning have shown that deep learning models are capable of developing representations.**

However, deep learning model are limited in their ability to understand context and provide meaningful explanations. Many of the early methods for detecting false information used manual extraction of linguistic attributes in conjunction with conventional statistical classifications. We are developing an end-to-end system for detecting fake news. This system combines a fine-tuned version of BERT (Bidirectional Encoder Representations from Transformers), with an Explainable AI Module.

The proposed system will be applied to the task of binary classification on social media posts and/or online news articles written in English that range from short-to-medium length. We have used WELFake as our main dataset for this study, and we intend to validate our results by comparing them with ISOT. We will also compare our results with those from other fake news datasets such as LIAR. Our methodology involves the following steps: cleaning and normalizing the data; tokenizing the data with BERT WordPiece tokenizer; creating sequences; fine-tuning BERT-base-uncased; performing binary prediction; providing post-hoc explanations of the predictions using LIME and SHAP; and visually presenting token-level importance. As this is a planning document for a future research study, we do not present actual accuracy, precision, recall, F1 score, latency, or statistically significant differences. Instead, we describe the methodology, evaluation protocol, system architecture, and expected outcomes that will guide the empirical evaluation of our proposed framework.

***Index Terms*—BERT, Explainable AI, fake news detection, LIME, misinformation, Natural Language Processing, SHAP, text classification, Transformer**

I. INTRODUCTION

The rise of social networking sites (SNSs), Internet-based news aggregators (i.e., news portals), and short-message posting services (microblogs) has changed how information is produced, disseminated, and consumed. Using these channels, an author's message can be viewed by many individuals instantly; this process does not require the traditional editorial review processes used in mainstream media. However, the same technologies that allow authors to quickly disseminate their messages to a wide number of people also enable authors to share unchecked, inaccurate, and intentionally deceptive information. In this summary, "fake news" refers to news reports/articles/posts that are demonstrably and provably false, and have been written/distributed with the intent of deceiving the reader for political, economic, or ideological reasons. [2]

In addition to the large quantity of data, there exist additional problems. False articles may simulate legitimate journalism and may also evoke strong emotional responses in readers. As mentioned in the reports, the fast spread of false information is due in large part to the novelty and emotional aspects of deceptive stories. Algorithms used by social media platforms to promote user interaction have created an environment where users are more likely to see exaggerated or extremely partisan content. Exposure to information which reinforces current beliefs within echo chambers and filter bubbles reduces the chances of users questioning either the source or accuracy of a statement.

Automated textual analysis is therefore a significant area of research given these circumstances. Fake-news detection systems traditionally have used manually created features including n-grams, word frequency counts, sentiment metrics, syntactic patterns, and TF-IDF to classify using SVM, Naive Bayes, or Logistic Regression.

While these methods may be sufficient for some types of controlled environments, they are typically limited in their ability to represent long-term semantic relationships and changes in context. [1]

Deep Learning Techniques are being developed with the use of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) Networks, which will decrease the amount of human time spent creating features. Instead of relying on hand-crafted features, these models develop a dense representation of words and phrases by looking at them in relation to their neighbors.

However, while CNNs and LSTMs may be able to identify local patterns and sequential dependencies, both of these methods still have limitations as to how much a particular word's meaning can change based on the words around it.

Transformer architectures provide a different method of representation through self-attention. Because BERT has bidirectional contextual representations and can look at the left and right context of each token over an entire encoded sequence, BERT will be useful when considering how deceptive language, inconsistencies in context and other subtle aspects of communication affect the understanding of information.

Therefore, this proposal suggests that BERT should be fine-tuned for binary fake news classification.

Even though a classifier has high predictive power it does not mean that the system is trustworthy. BERT-based classifiers make very complicated decisions which cannot be understood by a person using them (e.g., journalist, moderator, etc.). This research will address the "black-box" nature of BERT-based classifiers by applying LIME and SHAP as part of the process after classification. Both LIME and SHAP have been developed to help users understand what elements of the input (words/text) were most influential in making a specific prediction and to visualize these influences in an easy-to-read format.

The primary objective of the research is to provide a BERT-plus -XAI-based methodology that will examine both the classification results and the explanation results of an Artificial Intelligence (AI) model at the same time. The authors of the synopsis are especially interested in determining whether or not the explanations produced by their AI model are still relevant and consistent when the model has been able to achieve high levels of prediction accuracy and fast enough response times to support real-world use cases on social media platforms.

II. BACKGROUND AND RELATED CONCEPTS

2.1 Fake News and Digital Misinformation

This research defines "fake news" as intentionally misleading written material, rather than all forms of misleading information. This definition is important because misinformation may arise from errors or misinterpretations however, the focus of this study will be on content that is designed to deceive. Digital distribution allows a false statement to be seen by many people prior to a human being checking its accuracy. The results from this study highlight several different types of areas that would be significantly impacted by false information without proper verification. Examples of these areas include political manipulation, false information about public health, financial fraud and social conflict.

2.2 Evolution of Automated Detection

Computational problem development for detecting issues at the surface level started as a result of using surface-level feature engineering to produce context representation learning. The use of n-gram and tf-idf techniques enables the creation of document representation via lexical statistics. This is a relatively efficient method; however, it does not naturally represent the meaning of a word in relation to the surrounding context. CNNs can be used to recognize localized textual patterns. RNNs and LSTMs can be utilized to model sequential relationships.

Although these advancements reduced the reliance upon hand-crafted features, they still left unresolved the representation capabilities of static embeddings.

2.3 Transformer-Based Language Modelling

The transformer architecture utilized self-attention to relate tokens to each other, but did not process them in

order of their occurrence in a sequence (i.e., left-to-right). BERT took this idea further by creating a bidirectionally contextually encoded representation of the input and then performing a very large amount of pre-training. We plan to use the bert-base-uncased pre-trained model and fine tune it for binary fake news classification. The abstract describes the following characteristics of the model we will be using: an encoder with 12 layers, 768 hidden units, 12 attention heads, and around 110 million parameters.

2.4 Explainable Artificial Intelligence

Explainable Artificial Intelligence is relevant when a prediction needs to be inspected rather than merely accepted. LIME explains an individual prediction by generating perturbed versions of an input and fitting an interpretable local surrogate model. SHAP uses a Shapley-value formulation to assign contribution scores to features. In the proposed framework, both approaches are used to expose token-level evidence associated with a BERT decision. [9], [10]

III. LITERATURE REVIEW

The literature reviewed in this abstract is divided into four broad categories based on their utilization of methods or technologies; traditional machine learning, deep neural networks (DNN), Transformer Models (TM), and Post Hoc Explainability. Ahmed et al., applied n-gram representation to the problem of detecting online fake news via machine learning models such as SVM, SGD and Logistic Regression.

Ahmed, Traore, and Saad evaluated online fake-news detection using n-gram representations with machine-learning techniques including SVM, Stochastic Gradient Descent, and Logistic Regression. Their work demonstrates utility of lexical statistics and TF-IDF-style representations under controlled conditions; however, their work also illustrates limitation that manually engineered representations struggle to capture long-range semantic context. [1]

Allcott and Gentzkow examined the economic and social mechanisms surrounding fake news during the 2016 U.S. election cycle. Their work is important to the present research because it places fake news in a broader platform and social context rather than treating it only as a text-classification problem. [2]

Bondielli and Marcelloni surveyed fake-news and rumour-detection techniques and organized the area around content-based, context-based, and propagation-based approaches. Research utilizes this body of work, to motivate the progression from conventional textual features to deep-learning representations. [4]

Castillo, Mendoza, and Poblete studied information credibility on Twitter, demonstrating the importance of credibility analysis in microblogging environments. Their work also illustrates why social-media language cannot always be handled as conventional long-form news text. [5]

Cerisara, Kral, and Lenc examined the use of Word2Vec representations in neural dialogue-act recognition. Although the application domain is different from fake-news detection, the study is relevant to the synopsis discussion of static word embeddings and their role in earlier neural NLP systems. [6]

Devlin and colleagues introduced BERT as a pre-trained bidirectional Transformer language model. The contextual representation produced by BERT provides a stronger basis for text classification than static word vectors because the representation of a token depends on its surrounding context. [7]

Kaliyar, Goswami, and Narang proposed Fake-BERT, a BERT-based deep-learning approach for fake-news detection in social media. The study is directly relevant because it demonstrates the application of contextual Transformer representations to the target problem and motivates further investigation into how such models can be made interpretable. [8]

For explainability, Ribeiro, Singh, and Guestrin introduced LIME as a model-agnostic approach for explaining individual predictions through local surrogate models. Lundberg and Lee proposed SHAP as a unified feature-attribution framework grounded in Shapley values. These two methods provide complementary perspectives for examining the evidence used by an otherwise opaque classifier. [9], [10]

Table I. Comparison of fake-news detection and explainability approaches described in the synopsis.

Approach	Representative methods	Main strength	Main limitation for this study
Traditional ML	N-grams, TF-IDF, SVM, Logistic Regression	Efficient and suitable for smaller, controlled datasets	Limited semantic context and weak generalization to linguistic variation
Deep learning	CNN, RNN, LSTM / BiLSTM	Learns dense representations and local or sequential patterns	Static embeddings or sequential processing can limit contextual understanding
Transformer-based	BERT and related models	Bidirectional, context-aware representations	High computational cost and black-box decision process
Post-hoc XAI	LIME, SHAP	Provides human-readable feature attribution	Can add computational overhead and may approximate rather than directly expose internal reasoning

IV. RESEARCH GAP

The research overview has highlighted a disconnect between high performance predictive systems and the practical application of the interpretation of those systems' outputs.

While BERT-based classification tools successfully identify complex language environments, there is little clarity provided by BERT-based classification tools about why a user-related conclusion was reached. Since labels like "Real" or "Fake" are binary, they create an obstacle for fact-checkers and editors to determine what parts of a text caused a given result. A second disconnect relates to the relative weights of predictive accuracy and transparency. More basic architectures may be simpler to explain, but more basic architectures will likely be unable to discern the subtle intricacies of deceptive content within written communication. BERT, however, has a much deeper understanding of context than the more basic architectures, and therefore, provides greater system opacity.

While LIME and SHAP can be used to generate explanations of BERT's predictions, each of these methods adds significantly to the computational overhead associated with producing those explanations.

Therefore, the proposed research will consider accuracy, explanation quality, and computational cost as related research variables rather than separate areas of study.

Finally, there is a third gap concerning domain generalization. Because language varies depending upon the topic (e.g., politics, health), and because online communications often include slang, sarcasm, and highly partisan language, a model trained on one type of misinformation may not be able to perform as effectively on another type of misinformation. As such, the synopsis suggests using cross-domain evaluations as an operational objective.

V. PROBLEM STATEMENT

Research problem is to develop an automated system capable of classifying English-language news text or social-media posts as authentic or fake while concurrently providing comprehensible evidence for the classification. Let $D = \{(X_1,Y_1), \ldots, (X_n,y_n)\}$ denote a labelled textual dataset consisting of N samples. Each x_i is a token with maximum length T, and $y_i \in \{0,1\}$, where $y_i = 0$ denotes authentic/real news and $y_i = 1$ denotes fake/deceptive content.

The proposed classifier is represented by a parameterized function $f\ \theta: X \rightarrow [0,1]$, where θ denotes the learned BERT parameters.

The training objective specified in the synopsis is binary cross-entropy:

$$L_BCE(\theta) = -(1/N)\ \Sigma_i\ [\ y_i\ log(f\theta(x_i)) + (1-y_i)\ log(1-f\theta(x_i))\ ] \quad (1)$$

For explanation, the classifier f is treated as a black box and an interpretable model g is fitted around an individual input. For LIME, the objective is to minimize the discrepancy between f and g in a local neighbourhood while controlling surrogate complexity: $\arg\min_g [L(f,g,\pi_x) + \Omega(g)]$. For SHAP, the contribution $\varphi_j(x)$ of token j is expressed using Shapley-value weighting over feature subsets. These formulations are used as the conceptual basis for token-level attribution in the proposed framework.

The research consequently asks four central questions: (1) what trade-off exists between BERT detection accuracy and explainability for social-media text; (2) how much fine-tuning on domain-specific fake-news corpora improves F1-score relative to LSTM and CNN baselines; (3) how reliably LIME and SHAP identify deceptive linguistic structures, emotional triggers, and unsupported assertions; and (4) whether explainability can be incorporated into an inference pipeline without excessive latency for large-scale monitoring.

VI. RESEARCH OBJECTIVES

6.1 Primary Objectives

- Construct BERT-based Transformer model capable of performing binary (fake vs. real) news. classification for short-form social media text, and long-form text. news articles.
- Incorporate LIME and SHAP into the predictive inference pipeline to provide token-level dynamic attribution scores.
- Weigh up the trade-offs Between computational costs for explanations and predictive performance fiedity.
- Bench mark the BERT-XAI framework against SVM, Logistic Regression, CNN, and BiLSTM models using the publically available data sets referenced in the Synopsis.

6.2 Secondary and Operational Objectives

- Construct a preprocessing pipeline for removing unnecessary noise but preserving potentially useful structural cues (e.g., capitalization) that are relevant to the task.
- Develop an intuitive visualization tool which provides explanations with respect to humans by highlighting tokens with their importance and generating importance heat maps
- Test the stability of the generated explanations under different datasets and domains (i.e., politics to healthcare), in order to see if the explanation remains consistent over time.

VII. PROPOSED METHODOLOGY

Proposed method comprises multiple stages or "modules" that can be employed to collect and process train and infer from a Transformer model; generate explanations of the output produced; and visually present those generated explanations. The modular nature of this pipeline allows for comparative studies of different Transformer models (e.g., RoBERTa, DistilBERT) without altering the general concept behind how the pipeline works.

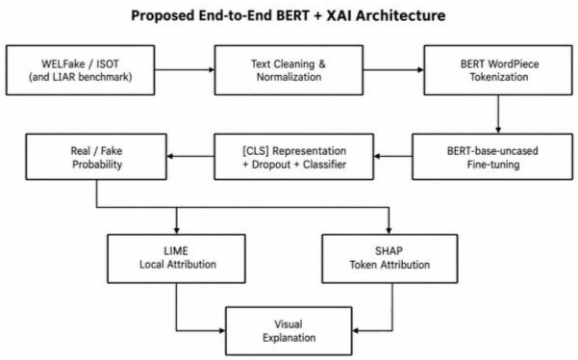


Fig. 1. Proposed end-to-end BERT + XAI system architecture.

7.1 Data Collection

The most important dataset referenced from reports is WELFake which has over 72,000 articles (both real and fake) combined. The next dataset to be used for cross-validation will be ISOT Fake News Dataset. LIAR will also be used in the standardized benchmarking scope. Since no specific train/validation/test split was provided in the synopsis, we did not provide a numerical split which was not included in the research proposal. Using multiple datasets are necessary to test how robust a model can perform. If a model only performs well on one corpus, it may have learned some type of vocabulary or source pattern from that specific corpus and not general indicators of deception.

Thus cross-dataset evaluation will be included as part of the proposed validation strategy.

7.2 Data Preprocessing

Raw text is expected to contain web formatting, URLs, HTML elements, punctuation variation, Unicode inconsistencies, and other structural noise. The proposed preprocessing stage performs normalization and cleaning while attempting to preserve useful semantic and stylistic signals. Lowercasing is proposed with selective retention of all-capitalization ratios because capitalization can itself be a stylistic indicator in deceptive online content.

- Remove HTML tags and irrelevant web metadata.
- Handle URLs and punctuation consistently.

- Apply Unicode normalization.
- Retain selected structural information, particularly capitalization-related signals where relevant.
- Prepare cleaned text for BERT WordPiece tokenization.
- Pad or truncate sequences to the proposed maximum length of 512 tokens for batch processing.

7.3 BERT Tokenization and Sequence Construction

The proposed model employs the typical BERT WordPiece tokenizer. The use of sub-word tokenization has been shown to be especially helpful when working with on-line language; this type of tokenization allows for representation of words that have not been seen before, or words which are infrequently used, by breaking them down into smaller pieces. Following tokenization, the [CLS] token will be added to the front of every input sequence and the [SEP] token will be added to the rear of every input sequence to designate the conclusion of the input sequence. Each input sequence will either be truncated or extended (padded) to a consistent maximum length. The maximum length of 512 tokens was suggested in the synopsis as one possible example of a configuration.

7.4 Transformer Fine-Tuning

The base model is BERT-Base-Uncased. According to the synopsis, this checkpoint contains 12 Transformer encoder layers, 768 hidden units per layer, 12 attention heads, and approximately 110 million parameters. It was pre-trained on English Wikipedia and Book Corpus. Fine-tuning adapts the general language representation to the binary fake-news classification task. The hidden representation associated with [CLS] is used as the aggregate sequence representation. A dropout layer with $p = 0.3$ is proposed to reduce overfitting, followed by a fully connected classification layer producing binary probabilities. The synopsis specifies AdamW as the optimizer, learning-rate evaluation in the range 2×10^{-5} to 5×10^{-5}, dynamic batch sizes of 16 or 32 depending on available VRAM, and 3–5 training epochs with early stopping based on validation loss.

These values are research-plan parameters rather than evidence of a completed experiment. The final empirical study should document the selected hyperparameters and justify any deviations from the proposed configuration.

7.5 Explainability with LIME

LIME uses local explanations of individual predictions by creating perturbed versions of an input example and observing how the classifier's output changes. The proposed approach creates synthetic variants of the target article by removing or masking words or tokens. The fine-tuned BERT model will predict each of these variants and the interpretable surrogate model will be fit around the original example. As stated in the synopsis, ridge regression will serve as the local surrogate and cosine similarity will serve as the weighting mechanism for the perturbed text.

The intended output is a ranked set of local feature contributions. Positive contributions can be presented as evidence supporting the predicted class, while negative contributions can indicate text that moves the prediction toward the opposite class. The final interface is expected to use readable highlights rather than expose the internal mathematical complexity of BERT.

7.6 Explainability with SHAP

An alternative approach for assigning credit involves utilizing Shapley values through the SHAP architecture. The existing plan intends to adapt a SHAP explainer designed for text to assess how much every individual token influences the final output probability. Conceptually, the importance assigned to each token reflects the shift it generates when integrated into different feature combinations.

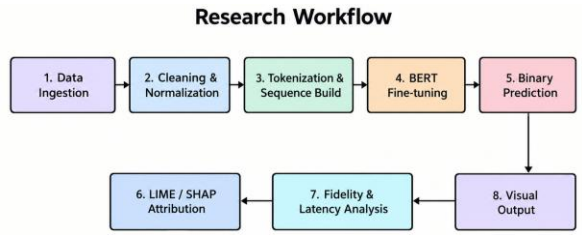
The summary identifies SHAP as a theoretical basis of comparison to LIME's use of a local surrogate approximation. Since the number of potential feature subsets increases exponentially with the increase in the number of tokens, there may be significant computational costs associated with applying SHAP to longer sequences. This issue has been identified as part of the research gap, and thus will be evaluated as an independent variable, rather than being excluded from consideration.

7.7 Visual Explanation Layer

The final stage converts attribution scores into an interface intended for non-specialist users. The synopsis proposes color-coded importance heatmaps, with deceptive or strongly influential fragments highlighted separately from authentic or context-supporting fragments.

The interface should preserve the original text and make the attribution understandable without requiring users to inspect model internals.

7.8 Research Workflow



VIII. SYSTEM ARCHITECTURE AND ALGORITHMIC WORKFLOW

The system consists of four interconnected modules. The first module accepts the benchmark corpora for normalization and tokenization. The second module contains the fine-tuned BERT classifier. The third module takes in both the input text as well as the classifier outputs to call LIME and SHAP to determine which features have an impact on the prediction.

Finally, the fourth module will display the predicted class with explanations in a format that is understandable to humans.

- Import the designated benchmark data set for use in this study; define the "truth" labels (ground truth) from the benchmark data.
- Remove noise (clean) and adjust formatting (normalize) of each text document while retaining some structural elements.
- Tokenize each text document using BERT's WordPiece algorithm. Construct a [CLS]-text-[SEP] sequence from these tokens.
- Adjust the length of each tokenized sequence so that all sequences are of equal length; either by adding padding or truncating as necessary.
- Send a batch of these adjusted tokenized sequences through the fine-tuned BERT model encoder and classification head.
- Determine the probabilities of both the "real" class and the "fake" class.
- If a user requests an explanation, create LIME perturbations and/or SHAP token attributions.
- Calculate aggregated attribution scores and convert them back into readable tokens/text snippets.
- Present the calculated classification along with the generated explanation. Record any relevant evaluation metrics including inference time when requested.

IX. EXPERIMENTAL SETUP

The plan for this experiment is to test all of these models (the main BERT-base-uncased model as well as those used in comparison) on the same platform where possible. The two primary comparisons to BERT are SVM and Logistic Regression which represent traditional machine learning, and CNN and BiLSTM representing deep learning. The WELFake corpus will be the focus of the evaluation while ISOT and LIAR will provide additional data for the larger benchmarking effort.

Table II. Proposed experimental configuration derived from the research synopsis.

Component	Proposed configuration from synopsis
Primary model	BERT-base-uncased
Encoder	12 Transformer layers; 768 hidden units; 12 attention heads
Approximate parameters	110 million
Classification representation	[CLS] hidden state
Dropout	p = 0.3
Optimizer	AdamW
Learning-rate search	2×10^{-5} to 5×10^{-5}
Batch size	16 / 32 depending on VRAM
Epochs	3–5 with early stopping
Maximum sequence length	512 tokens
Explainability	LIME and SHAP
Primary corpus	WELFake
Cross-validation / benchmark corpora	ISOT and LIAR

X. PERFORMANCE EVALUATION

The synopsis indicates that the model evaluation process will utilize all elements of the confusion matrix (TP (true positive), TN (true negative), FP (false positive), FN (false negative)) to assess model performance. The key classification metrics to be utilized in evaluating model performance include:

accuracy, precision, recall, and F1-score. These metrics must be reported for each dataset and model evaluated. This differs from evaluating improvements based on only one metric.

Accuracy represents the percentage of correctly classified instances. Precision indicates the frequency with which items identified as "fake" are indeed "fake"; recall indicates the effectiveness of the system in identifying actual "fake" instances.

The F1-score is a weighted average of precision and recall using the harmonic mean and is noted in the synopsis as being particularly useful when there exists a significant imbalance in class distribution.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

10.1 Explainability Evaluation

When evaluating the usefulness of an explanation it is not enough to simply look at classification metrics.

Therefore, this paper suggests using explanation stability as a metric to measure the effectiveness of explanations, particularly in relation to LIME. One way to do this would be to run the exact same input through LIME multiple times and see if the same significant tokens are highlighted each time.

Additionally, SHAP and LIME outputs could be qualitatively compared to determine if they agree on which phrases influence the output of the model.

In addition to measuring explanation quality and computational cost, the researcher should also be able to quantify the increase in processing time associated with generating explanations. This will allow for a quantitative assessment of the balance between accuracy and explainability. The goal of this research is not to suggest that any particular XAI method is universally better than another; rather, it is to evaluate how explanation quality and computational cost interact with each other in the context of the proposed fake-news detection pipeline.

10.2 Cross-Domain Robustness

The cross-domain evaluation is needed to determine if the models' ability to classify/explain behavior across domains (political news vs. health misinformation etc.) can be generalized.

Therefore, the proposed evaluation should assess classification and explanation behavior across all datasets included in this synopsis, as well as record how performance varies between these datasets, rather than making an assumption that the model that was developed to classify/explain behavior in one domain would automatically apply to other domains.

XI. RESULTS AND DISCUSSION

The synopsis that was submitted to us has stated the expected outcomes of the research and the proposed evaluation protocol. However, as the synopsis is based on the researchers' expectations about what their experiments will show and they have not yet measured anything in an experiment (i.e., we do not have any experimental data from them), we did not include any information regarding accuracy, precision, recall, F1 score, training time, explanation latency, or statistical significance within this section. We discuss below how the ultimate results of the research can be interpreted.

If the proposed BERT model performs better than the SVM, Logistic Regression, CNN, and BiLSTM baselines under the same evaluation conditions, it would provide evidence in favor of the claim that contextual Transformer representations are advantageous for the specified classification task.

The comparative analysis of XAI should assess if LIME and SHAP have identified some of the same linguistic evidence. If there are consistently highlighted meaningful deceitful phrases in both approaches, then this provides more qualitative support for the comparative analysis than a single explanation produced by a technique. Conversely, large differences in the results from each approach would indicate much about the reliability of the explanations and therefore need to be reported as opposed to being concealed.

Latency is an important measure of performance as well. Our proposed system will be used to monitor social media and in this context the generation of explanations will likely be much more costly than classification. As such, the final reported results should show how much classifier inference time there was compared to explanation time and how much additional cost was incurred as a result of using LIME/SHAP.

A decline in the model's performance as a result of domain shift is an additional factor that needs to be

considered during the final empirical discussion. If there is a decrease in the model's accuracy when applied to different topics or corpora, this does not mean that the model has failed; instead, it indicates a lack of ability to generalize across domains. This will provide motivation for future research using domain adaptation techniques.

XII. PROPOSED COMPARATIVE ANALYSIS

Table III. Comparative evaluation framework proposed by the synopsis.

Model family	Model / technique	Expected role in study	Main comparison dimension
Traditional ML	SVM	Baseline classifier	Accuracy, precision, recall, F1, computational cost
Traditional ML	Logistic Regression	Baseline classifier	Accuracy, precision, recall, F1, computational cost
Deep learning	CNN	Neural baseline	Contextual classification performance
Deep learning	BiLSTM	Sequential baseline	Long-range/sequential modelling comparison
Transformer	BERT-base-uncased	Primary proposed classifier	Classification performance and generalization
Transformer + XAI	BERT + LIME	Local explanation pipeline	Attribution stability and explanation latency
Transformer + XAI	BERT + SHAP	Attribution comparison	Attribution consistency and computational cost

XIII. EXPECTED CONTRIBUTIONS AND OUTCOMES

The main goal of the project is to create a fully optimized BERT-based fake-news classification system that can process english-language texts as described in the scope. The project's summary also states that it will improve upon both the precision and recall of the given traditional and deep learning baseline models, especially in regards to complex or highly structured deceptive materials. These are still considered potential outcomes until they have been verified through testing.

The second expected contribution is an integrated explainable AI (XAI) framework. In addition to creating a simple binary prediction from LIME and SHAP in combination with the BERT inference pipeline, the researchers' goal is to generate human-readable token-level evidence. The design of this dashboard/interface will visually display information based on color-coded importance heatmaps.

The third expected outcome is an empirical description of the accuracy-explainability trade-off. The study's goal is to quantify how long of an additional delay is caused by generating explanations and if the explanations generated match in a meaningful way with the model’s predictive behavior.

The fourth anticipated contribution is a reproducible experimental framework with a modular Python implementation that uses some of the same technology as stated in the synopsis (Hugging Face Transformers, PyTorch, and XAI). The synopsis suggests that the code base should also have good documentation to support its use and potential release as open source.

XIV. LIMITATIONS

- The proposed scope is restricted to English-language textual data and does not establish performance for multilingual or multimodal misinformation.
- The research focuses on textual content and therefore does not directly analyse images, videos, network propagation, or source credibility signals beyond what may be encoded in the text.
- The proposed BERT architecture has substantial computational requirements compared with conventional classifiers.

- Post-hoc explainability methods may approximate model behaviour rather than expose the complete internal reasoning of a Transformer.
- SHAP-style attribution can become computationally expensive for long token sequences.
- Cross-domain generalization remains uncertain until the proposed evaluation is completed.
- The synopsis does not specify final dataset splits, hardware, random seeds, or completed empirical results; these must be documented during implementation rather than inferred.

XV. FUTURE SCOPE

The proposed framework can be extended in several directions that follow directly from the identified limitations. Alternative Transformer backbones such as RoBERTa and DistilBERT can be evaluated within the modular architecture. Domain-adaptive fine-tuning can be investigated to improve robustness when moving between political, health, and other misinformation topics. Explanation generation can also be optimized for lower latency so that interpretability remains practical in high-volume monitoring environments.

Further research may examine multilingual fake-news detection, multimodal misinformation involving text and images, and the integration of contextual or propagation-level evidence. Such extensions should be treated as subsequent research directions rather than as capabilities claimed by the current proposed system.

XVI. CONCLUSION

This paper expands the submitted dissertation synopsis into a complete research framework for fake-news detection using a fine-tuned BERT language model with Explainable AI. The central research problem is the need to combine reliable binary classification with understandable evidence for the classification decision. The proposed architecture addresses this problem by using BERT for contextual text representation and LIME and SHAP for post-hoc feature attribution.

The methodology covers dataset acquisition, text normalization, WordPiece tokenization, BERT fine-tuning, binary prediction, explanation generation, visual interpretation, and evaluation of both classification and explanation behaviour. WELFake is identified as the primary dataset, with ISOT and LIAR included in the broader benchmarking strategy. SVM, Logistic Regression, CNN, and BiLSTM form the proposed comparison baselines.

A key feature of the research design is that it does not treat accuracy as the only success criterion. The study explicitly considers explanation stability, faithfulness, cross-domain robustness, and computational overhead. Because the synopsis contains expected rather than completed results, no numerical performance claims have been introduced into this paper. The eventual experimental phase should populate the proposed evaluation tables and establish whether the expected improvements are supported by evidence. Overall, the proposed BERT-plus-XAI framework provides a coherent basis for investigating the practical balance between predictive capability and transparency in automated fake-news detection. Its principal research value lies in evaluating these dimensions together within a single end-to-end pipeline.

ACADEMIC INTEGRITY NOTE

This research paper was expanded from the submitted dissertation synopsis. The synopsis remains the controlling source for the research problem, scope, proposed datasets, architecture, hyperparameters, evaluation strategy, expected outcomes, and listed references. Where the synopsis did not provide completed experimental measurements or implementation details, this paper deliberately avoids inventing them.

The additional Transformer reference by Vaswani et al. is included because the synopsis explicitly discusses the Transformer architecture while not listing the corresponding bibliographic entry in its reference section. The Asparouhov and Muthén reference supplied in the synopsis is retained for source fidelity, although it is not used as a substantive foundation for the fake-news methodology.

REFERENCES

[1] H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using N-gram analysis and machine learning techniques,” in *International Conference on Intelligent, Secure, and*

Dependable Systems in Distributed and Cloud Environments. Cham: Springer, 2017, pp. 127–138.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] T. Asparouhov and B. Muthén, "Weighted least squares estimation with missing data," *Mplus Technical Appendix*, pp. 1–10, 2010.

[4] A. Bondielli and F. Marcelloni, "A survey on fake news and rumor detection techniques," *Information Sciences*, vol. 497, pp. 38–55, 2019.

[5] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th International Conference on World Wide Web*, 2011, pp. 675–684.

[6] C. Cerisara, P. Kral, and L. Lenc, "On the effects of using word2vec representations in neural networks for dialogue act recognition," *Computer Speech & Language*, vol. 47, pp. 175–193, 2018.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[8] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11765–11788, 2021.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[11] A. Vaswani *et al*., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.