Explainable Inference for Hybrid Transfer Learning in Infodemic Misinformation Detection

Deepak Kanneganti†, Sajib Mistry†, Aneesh Krishna†, Mufti Mahmud‡, Monowar Bhuyan*
†School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Australia
‡ Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Saudi Arabia
*Department of Computing Science, Umeå University, Sweden
{s.kanneganti, sajib.mistry, a.krishna}@curtin.edu.au, mufti.mahmud@kfupm.edu.sa, monowar@cs.umu.se

***Abstract*—Infodemic events such as pandemics, wars, and elections accelerate the spread of misinformation on social media, demanding accurate and transparent detection models. Recent advances in hybrid transfer learning models improve detection performance by combining multiple representation stages. Despite their effectiveness, their multi-stage architecture lacks a unified prediction interface, preventing the effective use of standard explainable artificial intelligence (XAI) techniques. This lack of compatible inference mechanisms restricts model interpretability in high-stakes decision-making settings. This paper proposes an explainable hybrid transfer learning framework that addresses this inference gap. We introduce a probability-based predictive inference mechanism that enables plug-and-play integration of widely used post-hoc XAI techniques, including SHAP and ELI5, without modifying the underlying model architecture. Experiments on multiple real-world infodemic datasets show that the proposed approach outperforms state-of-the-art baselines while achieving a balanced trade-off between detection accuracy and explainability.**

***Index Terms*—Explainable AI (XAI), Hybrid Transfer Learning, Infodemic Misinformation, Fake News Detection.**

I. Introduction

Social media platforms such as Facebook, YouTube, and Twitter (X) have become major channels for information dissemination. During large-scale events such as the COVID-19 pandemic, wars, and elections, information spreads faster than verification mechanisms can respond [1]. The term *infodemic* refers to an excessive volume of information from multiple sources on social media about a specific event, where accurate information is mixed with rumors and false claims. Research shows that misinformation erodes public trust in government institutions and negatively affects economic conditions and public sentiment [2]. A study published in *Nature Medicine* reported that false claims during the COVID-19 infodemic reduced vaccination rates and increased vaccine hesitancy [3]. Examples include viral claims suggesting that vaccines could control individuals or that smokers were less vulnerable to COVID-19, leading to confusion and panic.

Misinformation detection is therefore critical for governments, social media platforms, and fact-checking organizations seeking to maintain transparency and public trust [4]. However, infodemic scenarios present unique challenges for automated detection systems. Information evolves rapidly, narratives shift frequently, and reliable labels are often delayed or incomplete. Under these conditions, traditional AI-based fake news detection methods struggle to adapt to changing content and emerging claims [5], [6]. The lack of adaptability in static detection models significantly limits their effectiveness in time-sensitive infodemic environments.

To address these challenges, transfer learning models based on large pre-trained language models such as BERT and BART have been widely adopted for misinformation detection [7], [8]. By leveraging contextual representations learned from large corpora, these models can be fine-tuned to domain-specific and low-resource settings, making them well suited to infodemic scenarios [9]. Recent research has further explored hybrid transfer learning architectures that combine multiple pre-trained models or learning stages to capture complementary linguistic and semantic features, resulting in improved robustness and predictive performance.

Despite these advances, hybrid transfer learning models introduce a critical limitation related to *explainability*. These models rely on multiple stages of processing, including tokenization, representation learning, and cross-model adaptation, which obscure the relationship between input features and final predictions. As a result, enabling local interpretation of model behaviour becomes significantly more challenging than for standalone transformer models [10]. This limitation restricts the practical adoption of hybrid models in high-stakes infodemic settings, where decision-makers require transparent and trustworthy explanations to support informed interventions.

Explainability is a crucial requirement for deploying misinformation detection systems in such contexts. Explainable artificial intelligence aims to transform black-box models into systems whose predictions can be understood by humans. Local explanation techniques such as Local Interpretable Model Agnostic Explanations, Shapley Additive Explanations, and Explain Like I'm Five have demonstrated effectiveness in interpreting neural networks and transformer-based classifiers [11], [12]. However, these techniques are primarily designed for single model architectures and cannot be directly applied to hybrid transfer learning models due to the absence of compatible inference mechanisms.

Existing studies on explainable misinformation detection

mainly focus on traditional natural language processing pipelines or standalone transformer models [6], [13], [14]. Although these works show that explanation techniques can provide valuable insights into model behaviour, they do not address the inference challenges posed by hybrid transfer learning architectures. Explainable systems such as exBERT [15], DOMAIN [16], and xFake [17] are similarly limited to individual transformer models. Consequently, there is currently no unified mechanism that enables widely used explainability techniques to operate effectively on hybrid transfer learning models for infodemic misinformation detection.

To address this gap, *we propose an explainable hybrid transfer learning framework that explicitly targets inference-level explainability*. The proposed approach introduces a unified inference mechanism that enables local interpretation of hybrid models without modifying their underlying architecture. We design two hybrid models based on BERT and BART to demonstrate how contextual and generative representations can be combined while preserving interpretability. The framework supports integration of explanation techniques such as SHAP and ELI5, allowing influential linguistic features to be identified at the prediction level. We evaluate the proposed framework through extensive experiments on multiple real-world datasets, including CovidAID [18], COVID19-FNIR [1], the LIAR dataset [19], and the Kaggle Fake News dataset [20]. The results demonstrate that the proposed models achieve superior detection performance compared to state-of-the-art baselines, while also enabling transparent and interpretable decision-making. The main contributions of this paper are summarized as follows:

- We propose an explainable hybrid transfer learning framework for infodemic misinformation detection.
- We introduce a unified inference mechanism that enables local interpretation of hybrid transfer learning models operating through multiple stages of transformation.
- We enable integration of widely used explainability techniques, including SHAP and ELI5, without modifying the underlying model architecture.
- We perform extensive experiments on real-world datasets to evaluate predictive performance, explainability, and computational efficiency.

II. Related Work

A. Infodemic misinformation detection methods

Prior research has extensively studied fake news detection by identifying misinformation sources and limiting their spread in online social networks [1], [2]. While effective for general analysis, these approaches are often not designed to handle the *time-sensitive and rapidly evolving nature of infodemic events*. Recent studies focus on infodemic detection using linguistic features and social media content to improve prediction accuracy [21], with Alonso et al. [22] demonstrating improved performance under high-volume information streams. Traditional NLP techniques, such as Bag-of-Words and TF-IDF [23], and dense embeddings like Word2Vec [24], have been widely used with classifiers including SVMs, Random Forests, and Naive Bayes [25]. Deep learning models, particularly RNNs, further enhance detection by capturing sequential dependencies in text [26]. Despite these advances, existing approaches largely prioritize detection performance and do not adequately address the *time-critical nature of infodemic misinformation*.

B. Transfer learning for misinformation detection

Transfer learning has become a key advancement in NLP, enabling pre-trained models to be fine-tuned for domain-specific tasks. Early work in computer vision shows that large-scale pre-training allows effective feature reuse across tasks [27], a principle later adopted in NLP through distributed representations such as Word2Vec and GloVe. The introduction of BERT marks a major breakthrough, leveraging bidirectional context modeling through feature-based and fine-tuning strategies [7], and achieving strong performance across benchmarks, including fake news detection. Jwa et al. [28] further demonstrate improved F1 scores using BERT-based fine-tuning. BART extends this paradigm by combining bidirectional encoding with auto-regressive generation [8], proving effective for tasks requiring both understanding and generation. Recent studies show that incorporating auxiliary sentence information during fine-tuning helps mitigate data sparsity and domain shift, improving misinformation classification performance [29]. Despite these advances, transfer learning models largely prioritize predictive accuracy, motivating the need for more interpretable approaches.

C. Explainable AI for misinformation detection

Explainability aims to enable human understanding of how a model makes decisions and has become increasingly important in domains that prioritize transparency and interpretability [30]. In misinformation detection, accuracy alone is insufficient, as infodemic events require both trustworthy predictions and clear explanations to support informed decision-making. Several explainable AI (XAI) frameworks have been developed to improve interpretability in such systems. For instance, DOMAIN provides explanatory insights by combining content categorization and source reliability analysis [16], while exBERT introduces an interactive visualization tool to support hypothesis formation about model reasoning [15]. The xFake system further advances explainability by incorporating multiple self-explainable components for attribute, semantic, and linguistic analysis, offering visual explanations to assist users in identifying misinformation [17]. Model-agnostic methods such as LIME and SHAP are also widely adopted in NLP to provide local and feature-level interpretations [11], [13]. However, existing XAI approaches are primarily designed for traditional NLP pipelines or standalone transformer models and do not address the inference and attribution challenges introduced by hybrid transfer learning architectures. Consequently, there is no unified mechanism that enables these techniques to operate effectively on hybrid transfer learning models in infodemic misinformation detection settings.

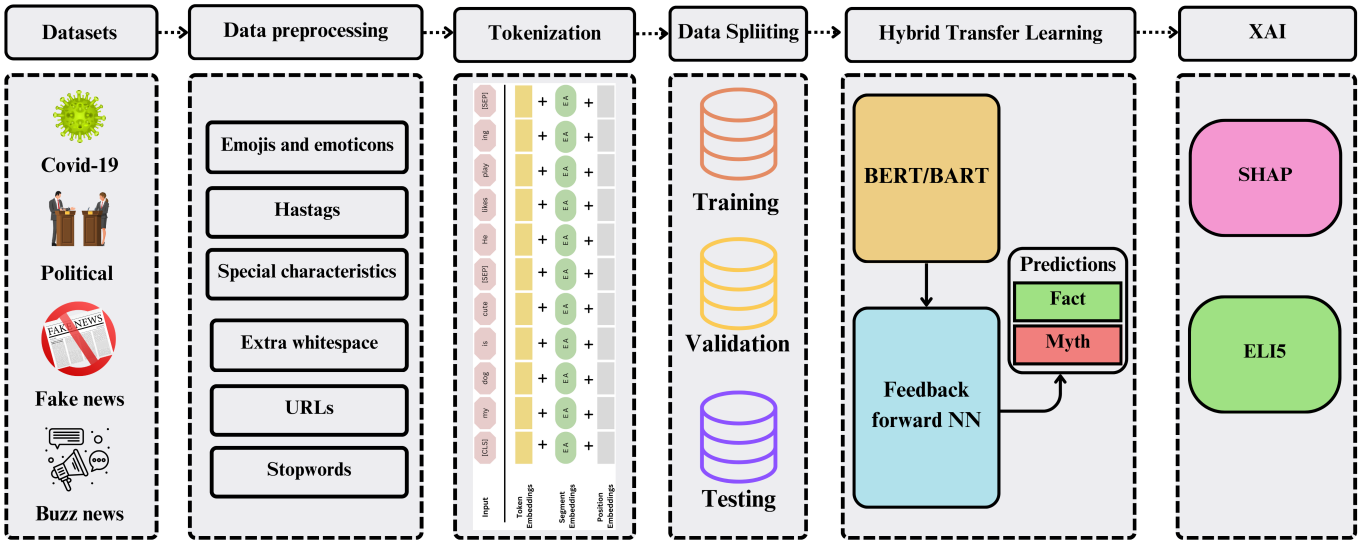

Fig. 1: Explainable hybrid transfer learning for infodemic misinformation detection

III. Explainable Inference in Hybrid Transfer Learning

The proposed hybrid transfer learning model for misinformation detection adopts a systematic design, as illustrated in *Fig. 1*. The model is evaluated using diverse infodemic datasets, such as COVID-19 updates, political news, fake news, and buzz news. The datasets undergo preprocessing and tokenization, followed by partitioning into training, validation, and testing sets to support robust evaluation. BERT and BART are employed as pre-trained transformer models to construct the hybrid transfer learning architecture. The model further incorporates Explainable AI (XAI) techniques, including SHAP and ELI5, to provide interpretability for model predictions.

A. Data preprocessing and tokenization

This study employs two pre-trained transfer learning models, BERT and BART, as the backbone of the proposed hybrid misinformation detection system. BERT is a bidirectional transformer encoder model consisting of 12 layers, 12 attention heads, and a hidden size of 768, pre-trained using Masked Language Modeling and Next Sentence Prediction to capture rich contextual representations. In contrast, BART is an encoder–decoder transformer architecture with 12 encoder and 12 decoder layers and a hidden size of 1,024, trained using a denoising autoencoder objective to support robust text reconstruction and generation. Leveraging the complementary strengths of these architectures enables effective handling of diverse linguistic patterns in infodemic data. For evaluation, widely adopted misinformation datasets are used, including CovidAID, COVID19-FNIR, LIAR, and Kaggle Fake News, primarily collected from social media platforms. The preprocessing pipeline removes emojis, URLs, special characters, extra whitespace, and stopwords, followed by normalization through lowercasing and punctuation removal to obtain consistent textual inputs. Tokenization follows the respective BERT and BART paradigms [7], [8], where special tokens `[CLS]` and `[SEP]` define sequence boundaries, inputs are padded or truncated to a fixed length of 512 tokens, and token-type IDs distinguish sentence pairs. BERT applies WordPiece tokenization, whereas BART employs byte-level BPE tokenization, both effectively managing out-of-vocabulary words by mapping subwords into integer identifiers for model processing.

B. SHAP-based hybrid transfer learning model architecture and training

This section describes the SHAP-based hybrid transfer learning model for infodemic misinformation detection, as summarized in *Algorithm 1*. The model takes a dataset of input texts D and a pre-trained transformer model m (BERT or BART) as inputs, and produces prediction probabilities together with SHAP-based explanations (Lines 1–4). The input data are first preprocessed to obtain a cleaned and standardized representation, followed by tokenization to generate input sequences and attention masks compatible with the selected model (Lines 5–7). The hybrid transfer learning architecture is then initialized using pre-trained parameters and trained iteratively across multiple epochs (Lines 8–10), where forward propagation generates predictions and backward propagation updates model parameters based on the computed loss (Lines 11–20). SHAP [31] is an explainable AI technique that interprets black-box models by attributing predictions to individual input features and providing local, instance-level explanations. In the proposed model, SHAP is integrated as a post-hoc explainability component through a probability-based predictive inference mechanism (Lines 21–22), operating over the multi-stage transformation pipeline, including tokenization, embedding generation, and intermediate representations. For each test instance, prediction probabilities are computed using the trained model, after which SHAP values are estimated to quantify feature-level contributions and generate explanation

scores, which are then visualized to support local interpretability of misinformation predictions (Lines 23–28).

Algorithm 1 SHAP-based Hybrid Transfer Learning Model for Infodemic Misinformation Detection ($\mathcal{H}$)

1: **Input:** Dataset $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$
2: **Model:** Pre-trained transformer model m (BERT/BART)
3: **Output:** Predicted label $\hat{y}$ and SHAP explanations $\boldsymbol{\phi}(x)$
4: Pre-process $D \rightarrow D_{\text{processed}}$
5: **function** TOKENIZE(x, m)
6: $(X_{\text{seq}}, X_{\text{mask}}) \leftarrow m_{\text{tokenize}}(x)$
7: **end function**
8: **Hybrid Transfer Learning Architecture:**
9: Initialize model parameters θ_m
10: **for** $epoch = 1$ to E **do**
11: **Forward Propagation:**
12: $h \leftarrow m(X_{\text{seq}}, X_{\text{mask}})$
13: Extract contextual representation h_{cls}
14: $z_1 \leftarrow f(h_{\text{cls}})$
15: $z_2 \leftarrow g(z_1)$
16: $p(x) \leftarrow \text{Softmax}(z_2)$
17: $\hat{y} \leftarrow \arg\max p(x)$
18: **Backward Propagation:**
19: Compute loss $\mathcal{L}(y_{\text{true}}, p(x))$
20: Update parameters θ_m
21: **end for**
22: **SHAP-based Explainability Integration:**
23: Initialize SHAP explainer $\mathcal{S}$
24: **for** each test instance $x^{(i)} \in D_{\text{test}}$ **do**
25: $p(x^{(i)}) \leftarrow$ PREDICT_PROBABILITY$(x^{(i)}, \mathcal{H})$
26: $\boldsymbol{\phi}(x^{(i)}) \leftarrow \mathcal{S}(x^{(i)})$
27: Compute model output score:

$$f(x^{(i)}) = \sum_{j=1}^{d} \phi_j(x^{(i)}) + \text{BaseValue}$$

28: Visualize SHAP values for $x^{(i)}$
29: **end for**

C. ELI5-based hybrid transfer learning model architecture and training

Using Explain Like I'm 5 (ELI5) [32], we design an ELI5-based hybrid transfer learning model that enables instance-level explanations within complex hybrid architectures. Prior studies show that ELI5 provides effective interpretations for several machine-learning algorithms by highlighting word-level weights and assigning positive and negative contribution scores. However, the architectural complexity of hybrid transfer learning models makes direct adaptation of such explainability techniques challenging due to multiple transformation stages. To address this, we introduce the pipeline function `predict_probability` (Algorithm 2, Lines 1–10) as a probability-based predictive inference mechanism that systematically connects the tokenizer output with the model function. This mechanism operates across tokenization, embedding generation, and intermediate representation learning to bridge internal computations with interpretable outputs. The TextExplainer uses 100 samples to estimate prediction probability and an expansion factor of 10 to generate explanations for each instance. The ELI5 explainer outputs the prediction probability, and the resulting score reflects the model's behaviour (Algorithm 2, Lines 10–17).

Algorithm 2 ELI5-based Explainable Hybrid Transfer Learning Model ($\mathcal{H}$)

1: **Input:** Dataset $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$
2: **Model:** Pre-trained transformer model m (BERT/BART)
3: **Output:** Prediction probability $p(x)$ and ELI5 explanations $\boldsymbol{\psi}(x)$
4: Pre-process $D \rightarrow D_{\text{processed}}$
5: **function** TOKENIZE(x, m)
6: $(X_{\text{seq}}, X_{\text{mask}}) \leftarrow m_{\text{tokenize}}(x)$
7: **end function**
8: **ELI5-based Explainability Integration:**
9: Initialize ELI5 explainer $\mathcal{E}$
10: **for** each test instance $x^{(i)} \in D_{\text{test}}$ **do**
11: $p(x^{(i)}) \leftarrow$ PREDICT_PROBABILITY$(x^{(i)}, \mathcal{H})$
12: $\boldsymbol{\psi}(x^{(i)}) \leftarrow \mathcal{E}(x^{(i)}, p(x^{(i)}))$
13: Compute explanation score:

$$g(x^{(i)}) = \sum_{j=1}^{d} \psi_j(x^{(i)})$$

14: Visualize feature importance for $x^{(i)}$
15: **end for**

IV. Experiments and Results

In this section, we evaluate our model on four widely studied datasets (e.g., [13], [14]). First, experiments were conducted to evaluate the predictive accuracy of the proposed hybrid BERT and BART models for misinformation detection. Second, we generated explanations from the hybrid BERT/BART models using surrogate XAI techniques across four representative classification scenarios. Third, we analyzed the computational efficiency of different XAI methods by measuring their explanation generation time to assess suitability for time-sensitive infodemic settings. All experiments were performed on a computer with an Intel Core i7 CPU with 16 GB RAM using Python. The experimental results and source code are available in our GitHub repository[1].

A. Data description

We evaluated the explainable hybrid transfer learning model using datasets derived from various infodemic events, including the COVID-19 pandemic, the US election, and other political events. The primary datasets used are CovidAID, COVID19-FNIR, the LIAR dataset, and the Kaggle Fake News dataset, all publicly available and collected from social media platforms such as Twitter and Facebook, as well as

[1]https://anonymous.4open.science/r/XAIExplainability-04AC

TABLE I: Performance comparison of HTLM and baseline approaches across multiple infodemic datasets.

Model	COVID AID			COVID-19 FNIR			Kaggle			LIAR		
	Acc	Pre	F1	Acc	Pre	F1	Acc	Pre	F1	Acc	Pre	F1
Hybrid BART	**0.97**	**0.97**	**0.97**	0.97	0.965	0.965	**0.97**	**0.973**	**0.975**	**0.653**	**0.652**	**0.642**
Hybrid BERT	0.96	0.97	0.965	**0.98**	**0.985**	**0.985**	0.93	0.934	0.935	0.63	0.63	0.62
BERT [13]	0.94	0.95	0.94	0.95	0.95	0.97	0.95	0.95	0.95	0.622	0.617	0.617
BART [8]	0.94	0.95	0.94	0.94	0.945	0.945	0.94	0.945	0.94	0.632	0.631	0.631
Bi-LSTM [14]	0.931	0.94	0.93	0.95	0.97	0.95	0.92	0.93	0.93	0.613	0.621	0.613

other online sources. The original CovidAID dataset contains over 10,000 COVID-19–related social media posts annotated as misinformation or factual content, while COVID19-FNIR includes approximately 20,000 samples focusing on false narratives related to COVID-19. The LIAR dataset provides over 12,800 manually labeled short statements from PolitiFact.com, and the Kaggle Fake News dataset comprises around 25,000 news articles labeled as fake or real. After preprocessing and label standardization, the datasets contain 28,100 instances for CovidAID, 10,700 for COVID19-FNIR, 12,800 for LIAR, and 35,600 for the Kaggle dataset, as summarized in Table II. These datasets provide a robust foundation for evaluating the performance and interpretability of the proposed hybrid transfer learning models.

TABLE II: Distribution of class labels across all datasets

Dataset	Fact Instances	Myth Instances	Total Instances
CovidAID	14,000	14,100	28,100
COVID19-FNIR	5,700	5,000	10,700
LIAR dataset	6,200	6,600	12,800
Kaggle dataset	18,000	17,600	35,600

B. Experiment 1: Performance comparison of hybrid transfer learning models for infodemic misinformation detection

In this section, we evaluate the performance of the proposed hybrid transfer learning models on multiple infodemic datasets, including CovidAID [18], COVID19-FNIR [1], the LIAR dataset [19], and Kaggle's Fake News dataset [20]. *Table I* summarizes the comparative performance across hybrid and baseline models, demonstrating that fine-tuned hybrid architectures consistently outperform standalone models across all datasets. On the COVID-19 AID dataset, Hybrid BART achieves the best performance, with an accuracy and F1-score of *0.97*, outperforming Hybrid BERT (accuracy 0.96, F1 0.965) and baseline BERT, BART, and Bi-LSTM models. While standard BERT and BART achieve accuracies around 0.94, and Bi-LSTM achieves 0.931, the hybrid approach clearly captures richer infodemic patterns through effective transfer learning. For the COVID-19 FNIR dataset, Hybrid BERT achieves the highest accuracy and F1-score of *0.98* and *0.985*, respectively, marginally outperforming Hybrid BART (accuracy 0.97, F1 0.965). Baseline BERT, BART, and Bi-LSTM models show slightly lower but comparable performance, reinforcing the advantage of hybrid fine-tuning for misinformation narratives. On the Kaggle dataset, Hybrid BART again achieves the highest performance with an accuracy of *0.97* and an F1-score of *0.975*, followed by Hybrid BERT with an accuracy of 0.93 and an F1-score of 0.935. In contrast, the Bi-LSTM model records a lower accuracy of 0.92, indicating reduced effectiveness compared to transformer-based hybrid models. The LIAR dataset remains the most challenging due to its domain specificity and class imbalance. Among all evaluated models, Hybrid BART achieves the highest accuracy of *0.653* with an F1-score of *0.642*, followed by Hybrid BERT with an accuracy of 0.63. Although performance on LIAR is lower overall, the hybrid models consistently outperform baseline approaches, highlighting their improved robustness in politically sensitive infodemic settings.

C. Experiment 2: Interpretability analysis using SHAP and ELI5-based hybrid transfer learning models

In this section, we present the interpretability analysis of the hybrid transfer learning model using SHAP and ELI5. The binary classification task distinguishes between factual information and misinformation (myth), where misinformation is treated as the positive class, resulting in four prediction scenarios: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Representative text samples and their corresponding outcomes are summarized in Table III to illustrate the model's local interpretability.

1) SHAP explanation for hybrid transfer learning: Explanations obtained using the SHAP model are analyzed based on the *base value*, $f(x)$, *highlighted words*, and the corresponding *weight values* of each observation. The base value $E(y')$ represents the mean model prediction in the absence of input features, while $f(x)$ denotes the prediction score for a specific instance. As shown in *Fig. 2*, the prediction is obtained by combining the base value with feature contributions, where SHAP values indicate how each word shifts the prediction toward or away from the final class. The SHAP visualization further highlights that words with **'+'** weights increase the model's confidence in the predicted class, while words with **'–'** weights reduce that confidence.

The test statement, *"In a community setting, remember the 3 Ws: Wash your hands..."*, is correctly classified as factual information. Words such as 'the', 'Wash', and 'distance' contribute toward the predicted outcome, while terms like 'spreading', 'In', and 'CO' have an opposing influence. The prediction score is $f(x) = 0.008$ with a base value of $E[f(X)] = 0.283$, supporting a confident factual outcome. As

TABLE III: Classification outcomes with representative text samples illustrating model prediction behavior.

Category	Label	Text Sample
True Positive (TP)	Myth	In a community setting remember the 3 Ws. Wash your hands Watch your distance stay 6 feet apart. Wear A Mask These habits can help you protect yourself and others from spreading COVID-19.
True Negative (TN)	Fact	A photo shows a 19 year old vaccine for canine coronavirus that could be used to prevent the new coronavirus causing COVID19.
False Positive (FP)	Fact	FA first vaccine while a welcome tool in fighting COVID-19 may turn out to be of limited use The lesson learned from AIDS is the value of building a scientific infrastructure beyond a vaccine.
False Negative (FN)	Myth	At this rate India is slated to overtake USA in testing.

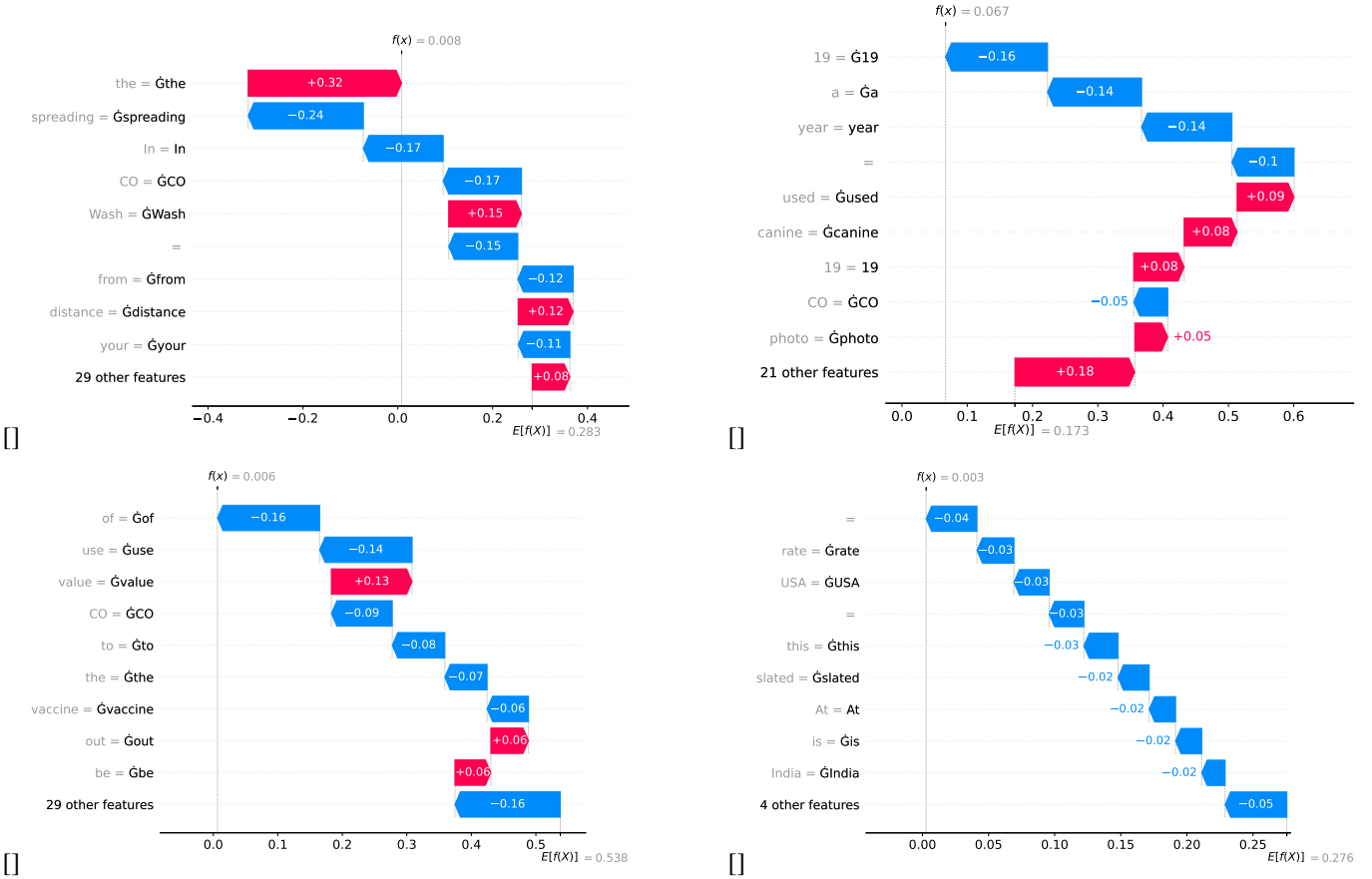

Fig. 2: SHAP waterfall explanations for four representative classification outcomes, illustrating feature-level contributions for true positive (a), true negative (b), false positive (c), and false negative (d) predictions.

shown in *Fig. 2(a)*, instance such as *"A photo shows a 19-year-old vaccine for canine coronavirus…"* are classified as myths, with the SHAP explainer correctly assigning a positive $f(x)$ value of 0.067 and a base value of 0.173. Words such as 'photo', 'canine', '19', 'used', and 'year' receive positive ('+') weights, contributing to myth predictions, as shown in *Fig. 2(b)*. The third instance, *"A first vaccine while a welcome tool in fighting COVID-19…"*, is misclassified as a fact, with an $f(x)$ value of 0.003 and a base value of 0.276, indicating insufficient learned patterns for accurate fact identification. Highlighted words such as 'India', 'USA', 'At', 'slated', 'this', and 'rate' suggest confusion in identifying relevant features, as evidenced by the SHAP explanation in *Fig. 2(c)*. Lastly, the model misclassifies the instance *"At this rate, India is…"* as a myth, where words such as 'India', 'rate', 'USA', 'slated', and 'At' dominate the explanation. The corresponding values ($f(x) = 0.003$, base value $E[f(X)] = 0.276$) indicate a low prediction score shifted from the baseline, reflecting uncertainty in the model's decision, as illustrated in *Fig. 2(d)*.

2) ELI5 explanation for hybrid transfer learning: The explanation generated using ELI5 is described in *Table IV*. The four test scenarios are presented along with the model's prediction probabilities and scores. The influence score represents the impact of different features, where both positive and negative contributions are shown. The first sentence *"In a community setting, remember the 3 Ws. Wash your hands…"* shows a high confidence that the information is factual. The ELI5 explainer produces a prediction close to 1, and the negative influence score is -14.22, indicating that most words support the classification as a fact. The second sentence *"A*

TABLE IV: ELI5 explanations of the hybrid-BART model

Cases	Text	Prediction	Probability	Score
True Positive	In a community setting, remember the 3 Ws. Wash your hands, watch your distance, stay 6 feet apart, wear a mask. These habits can help you protect yourself and others from spreading COVID-19.	Fact	1	-14.22
True Negative	A photo shows a 19-year-old vaccine for canine coronavirus that could be used to prevent the new coronavirus causing COVID-19.	Myth	1	5.31
False Positive	A first vaccine, while a welcome tool in fighting COVID-19, may turn out to be of limited use. The lesson learned from AIDS is the value of building a scientific infrastructure beyond a vaccine.	Myth	1	12.17
False Negative	At this rate, India is slated to overtake the USA in testing.	Fact	0.99	-5.55

photo shows a 19-year-old vaccine for canine coronavirus... " is an instance of a false positive. The ELI5 explainer shows high confidence in predicting it as a myth, with a probability of 0.995 and a positive influence score of 5.31. In the third test case, the model incorrectly classifies factual information as a myth. The probability provided by ELI5 for the instance *FA first vaccine while a welcome tool in fighting COVID-19...* shows high confidence; however, the model remains uncertain due to unfamiliar patterns, indicating that additional samples related to terms such as 'fighting', 'COVID-19', 'AIDS', and 'vaccine' are required to better distinguish myths from facts. Lastly, the model classifies the sentence *At this rate, India is...* as a myth. The analysis of the prediction probability and influence scores suggests that the decision is inaccurate, and additional training samples are required to improve the model's reliability.

D. Experiment 3: Comparative analysis of XAI approaches for explanation accuracy and efficiency in hybrid models.

In this experiment, three evaluation metrics were used to assess the performance of the explainable AI (XAI) models: accuracy, confidence, and computation time. Following the line-based evaluation strategy commonly adopted in i-flash [13], samples from all datasets were used to assess how effectively the proposed XAI methods enhance the explainability of the hybrid transfer learning model. Accuracy measures correct predictions, confidence reflects alignment between XAI explanations and hybrid model predictions, and computation time captures the efficiency of generating predictions and explanations. *Table V* presents the comparative performance of hybrid explainable models based on BERT and BART, where LIME-Hybrid BERT and SHAP-Hybrid BERT achieved the highest accuracy of 0.96 with confidence scores of 0.99, indicating strong predictive reliability for BERT-based models. ELI5-Hybrid BERT showed moderate performance with an accuracy and confidence of 0.84. For BART-based models, ELI5-Hybrid BART achieved high accuracy (0.95) with a confidence of 0.97, while LIME-Hybrid BART demonstrated both high accuracy (0.95) and the highest confidence (0.99), making it the most reliable overall. In contrast, SHAP-Hybrid BART recorded the lowest accuracy (0.79) and confidence (0.81), indicating comparatively weaker performance. Computation time was evaluated to assess the suitability of explainable models for time-sensitive and resource-constrained infodemic applications. As shown in *Fig. 3a,b*, LIME required the least computation time, with LIME-Hybrid BERT and LIME-Hybrid BART taking 413.53 seconds and 322.43 seconds, respectively, for 2000 instances. In contrast, SHAP-based models incurred substantially higher computation times, with SHAP-Hybrid BERT and SHAP-Hybrid BART requiring 1603.98 seconds and 1404.22 seconds due to the complexity of generating explanations. ELI5 exhibited intermediate performance, taking 915.74 seconds for Hybrid BERT and 455.17 seconds for Hybrid BART. Overall, LIME is most suitable for efficiency-critical scenarios, SHAP is appropriate when detailed interpretability is required despite higher cost, and ELI5 provides a balanced trade-off between explanation quality and efficiency.

TABLE V: Performance metrics of XAI models for hybrid transfer learning

Method	Accuracy	Computation Time (Sec)	Confidence
LIME–Hybrid BERT [13]	**0.96**	**413.53**	**0.99**
LIME–Hybrid BART [13]	0.95	322.43	0.99
SHAP–Hybrid BERT	**0.96**	1603.98	**0.99**
SHAP–Hybrid BART	0.79	1404.22	0.81
ELI5–Hybrid BERT	0.84	915.74	0.84
ELI5–Hybrid BART	**0.95**	455.17	**0.97**

V. Concluding Remarks

In this paper, we propose an explainable hybrid transfer learning model for infodemic misinformation detection. The proposed framework leverages hybrid BERT and hybrid BART architectures to address the complex and rapidly evolving nature of infodemic data, where contextual ambiguity and limited labeled information pose significant challenges. The hybrid models consistently demonstrate stronger detection performance than standard transfer learning and traditional baseline techniques across existing benchmark infodemic datasets. The hybrid models show a consistent accuracy gain of approximately *3–4%* on average over standard BERT, BART, and Bi-LSTM models. We integrate explainable AI techniques

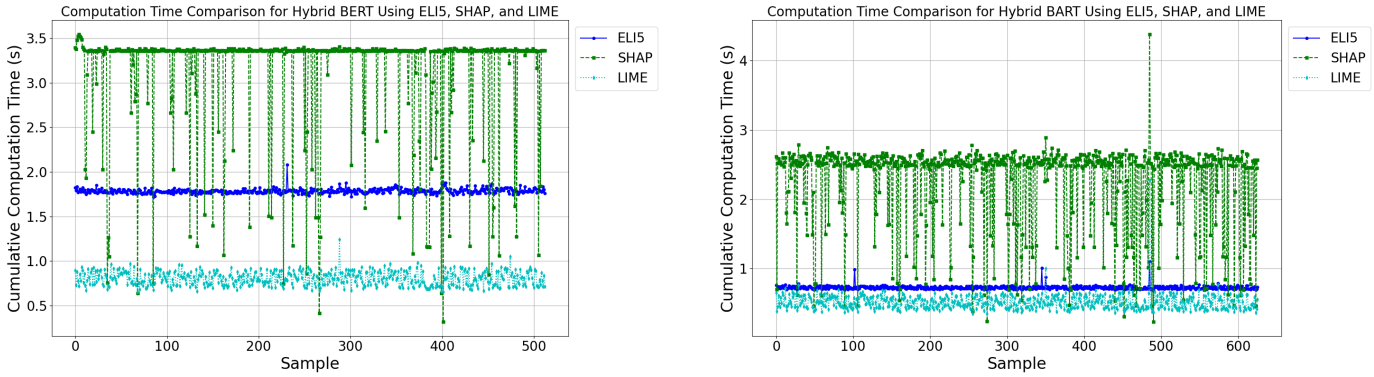

Fig. 3: Comparison of explanation computation time for hybrid BERT and hybrid BART using LIME, SHAP, and ELI5.

as plug-and-play modules by designing a predict–probability interface that enables SHAP and ELI5 to generate explanations for the proposed hybrid transfer learning models. We further evaluate the proposed explainable approach against widely used state-of-the-art explanation approaches, including LIME. The results show that LIME is more computationally efficient, whereas SHAP and ELI5 provide complementary explanation behavior with greater interpretive detail in selected cases, albeit at higher computational cost. Future work will focus on advancing explainability techniques by integrating cross-domain knowledge maps to enhance the model's adaptability in detecting misinformation. Leveraging interpretable analysis, we aim to address domain-specific data gaps and refine embedding techniques through knowledge distillation. Knowledge distillation enables the transfer of misinformation-specific patterns from a specialized teacher model to a streamlined student model, facilitating the use of explainability methods that ensure the student model's effective application in real-time scenarios.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] J. A. Saenz, S. R. K. Gopal, and D. Shukla, "Covid-19 fake news infodemic research dataset (covid19-fnir dataset)," *IEEE Dataport*, 2021.
[2] M. F. Mridha and et al, "A comprehensive review on fake news detection with deep learning," *IEEE access*, 2021.
[3] F. Pierri and et al., "Online misinformation is linked to early covid-19 vaccination hesitancy and refusal," *Scientific reports*, 2022.
[4] S. B. Naeem and et al, "The covid-19 'infodemic': a new front for information professionals," *HILJ*, 2020.
[5] S. Raponi and et al., "Fake news propagation: A review of epidemic models, datasets, and insights," *ACM TWEB*, 2022.
[6] D. Kanneganti, "A new cross-domain strategy based xai models for fake news detection," *arXiv preprint arXiv:2302.02122*, 2023.
[7] J. Devlin and et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*, 2019.
[8] M. Lewis and et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020.
[9] R. K. Kaliyar and et al., "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *MTA*, 2021.
[10] P. N. Ahmad and et al., "Misinformation detection on online social networks using pretrained language models," *IPM*, 2026.
[11] M. T. Ribeiro and et al., "" why should i trust you?" explaining the predictions of any classifier," in *ACM SIGKDD*, 2016.
[12] P. Gohel and et al., "Explainable ai: current status and future directions," *arXiv preprint arXiv:2107.07045*, 2021.
[13] V. Dua and et al., "I-flash: Interpretable fake news detector using lime and shap," *WPC*, 2023.
[14] M. Szczepański and et al., "New explainability method for bert-based model in fake news detection," *Scientific reports*, 2021.
[15] B. Hoover and et al., "exbert: A visual analysis tool to explore learned representations in transformer models," in *ACL*, 2020.
[16] D. Caled and et al., "Domain: Explainable credibility assessment tools for empowering online readers coping with misinformation," *ACM TWEB*, 2024.
[17] F. Yang and et al., "Xfake: Explainable fake news detector with visualizations," in *WWW*, 2019.
[18] C. Whitehouse and et al., "Evaluation of fake news detection with knowledge-enhanced language models," in *ICWSM*, 2022.
[19] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv*, 2017.
[20] Kaggle, "Fake news detection dataset," https://www.kaggle.com/c/fake-news/data?select=train.csv, 2024.
[21] M. La Morgia and et al., "Pretending to be a vip! characterization and detection of fake and clone channels on telegram," *ACM TWEB*, 2018.
[22] M. A. Alonso and et al., "Sentiment analysis for fake news detection." MDPI, 2021.
[23] S. Kaur and et al., "Automating fake news detection system using multi-level voting model," *Soft Computing*, 2020.
[24] T. Mikolov and et al., "Distributed representations of words and phrases and their compositionality," *NeurIPS*, 2013.
[25] M. A. Chandra and et al., "Survey on svm and their application in image classification," *IJIT*, 2021.
[26] A. Choudhary and et al., "Linguistic feature based learning model for fake news detection and classification," *ESA*, 2021.
[27] M. Biesialska and et al., "Continual lifelong learning in natural language processing: A survey," *arXiv*, 2020.
[28] H. Jwa and et al., "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert)." MDPI, 2019.
[29] S. Qin and et al., "Boosting generalization of fine-tuning bert for fake news detection," *IPM*, 2024.
[30] M. Szczepański and et al., "The methods and approaches of explainable artificial intelligence," in *ICCS*. Springer, 2021.
[31] S. M. Lundberg and et al., "A unified approach to interpreting model predictions," *NeurIPS*, 2017.
[32] A. Fan and et al., "Eli5: Long form question answering," *arXiv*, 2019.