Journal of Scientific & Industrial Research
Vol. 85, April 2026, pp. 325-340
DOI: 10.56042/jsir.v85i4.22599

BiSplit-DistilBERT: A Lightweight Early-Exit Transformer for Fake News Detection with Cross-Domain Evaluation on BoolQ Question-Answering Data Benchmarked against BERT, RoBERTa, and DeBERTa

Aradhana Saxena* & A. Santhanavijayan

Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli 620 015, Tamil Nadu, India

Received 10 August 2025; revised 27 February 2026; accepted 14 May 2026

In the age where people communicate more on digital mediums, authenticity of content is a benchmark. At the same time light weighted models are more preferable. By considering both things a light weighted model is developed in this study by modifying the classification layer of DistiBERT using Bi-Split method. The idea behind Bi-Split method is that prediction is possible by only the first half of the sentence in such cases. This generates an adaptive early-exit approach, in which the model determines whether to end prematurely or proceed with further processing to gain further insight. Upon reaching a predefined threshold, a prediction is made, otherwise the remaining text is analysed to maintain accuracy. The model is tested on four benchmark datasets, GossipCop, PolitiFact, ISOT, and BoolQ, using Accuracy, Precision, Recall, F1-score, and AUC. Accuracies of 99, 94.8, 91.5 and 85% are achieved, showing better performance than baseline models, with SHAP-based interpretability.

Keywords: Adaptive inference, Binary split classification, Explainable artificial intelligence, Fake news detection, Hierarchical classification

Introduction

The communication environment in the world has changed drastically due to the tremendous increased digital information, especially in an event like the COVID-19 pandemic. There has been a move towards fast online distribution with online sources becoming more and more dependable as information sources in areas like public health, education, governance, and finance. Sadly, this increased flow of information has also facilitated uncontrolled dissemination of misinformation which has resulted in skewed accounts, poor judgments, and loss of trust in the population. Since nowadays textual content can be produced not only by humans, but also by the language model with AI, the authenticity and verifiability of such content has become an increasingly prominent concern. Even though some methods of Machine Learning (ML), Deep Learning (DL), and transformer-based misinformation detection have been developed, the increasing amount and pace of data have become a challenge to address. In particular, lightweight models that can operate effectively and consume little computational resources and maintain accuracy are required. Fake news detection can be viewed as a text classification task, where unstructured textual input is assigned binary or multi-class veracity labels.[1–3] Notable progress in this domain has been achieved with transformer architectures like BERT[4] and its many variants[5], that use self-attention to achieve high performance in context-conditional classification. The large parameter sizes and deeply structured architectures of these models, however, make them typically not to be suitable in real-time or edge-based applications. In order to reduce these shortcomings, smaller versions like DistilBERT have been proposed in earlier research.[6] This simplified version maintains almost 97% the performance of BERT with 40% fewer parameters. Although efficient, these small models have rarely been tested in terms of their versatility across areas, especially when it comes to swapping between short fake news headlines to structurally varied models like fact-based question-answering datasets like BoolQ. As a result, a significant research gap has been established in the architecture design of systems that can provide resource efficiency, semantic adaptability, and explainability in a joint manner. To overcome this difficulty, a new paradigm that is called BiSplit-

*Author for Correspondence
E-mail: aradhana298@gmail.com

DistilBERT has been suggested in the current work. This architecture is based on the DistilBERT backbone, but it is augmented with a recursive binary splitting process in the classification head to enhance label discrimination. The model has been fine-tuned in two phases: Firstly with misinformation data sets (GossipCop, PolitiFact and ISOT) and then by domain-specific adaptation on BoolQ so that it can generalize to text formats. Experimental evaluations have shown promising results, with BiSplit-DistilBERT being competitive with larger transformer models. Accuracy scores of 94.8 on GossipCop, 91.5 on PolitiFact, 99 on ISOT and 85.0 on BoolQ have been achieved, and are all generally higher than BERT, RoBERTa and DeBERTa. These results confirm the ability of the suggested model to be able to maintain high performance whilst being computationally lightweight. With this integration, model transparency has been enhanced, balancing its design with ideas of ethical AI. This renders the model especially appropriate to be put into action in sensitive settings like media validation, government policy enforcement, and regulatory management. Even though previous approaches like RoBERTa-LightGBM[6], domain adversarial training[7], and multi-modal fake news detection systems[8] have demonstrated partial success, in many cases, they have obtained gains in one task at the expense of another. Conversely, BiSplit-DistilBERT model provides a comprehensive solution by enhancing computational efficiency, cross-domain robustness, and interpretability in a single framework. In this regard, it adds a scalable and realistic solution to reliable detection of misinformation in contemporary digital ecosystems.

The main contributions of this work can be summarized as follows:

- It is suggested to add a new lightweight model, BiSplit-DistilBERT, to DistilBERT by introducing a binary-split classification head to detect fake news efficiently.
- A dynamic early-exit mechanism is presented, which saves computation time and space by only processing the first half of the input text when confidence thresholds are met, thus enhancing computational efficiency by saving effective inference workload.
- A cross domain evaluation strategy is integrated, by using three fake news datasets (PolitiFact, GossipCop, and ISOT) and one factual question-answering dataset (BoolQ), which allows testing model generalization on varied textual formats.
- To increase model interpretability and improve the transparency of model predictions, which can be aligned with human reasoning, an explainable AI framework that uses SHAP is included.
- The following research questions have been used in organizing the present work:

RQ1: Why is DistilBERT appropriate to detect fake content with low-resource?

RQ2: How does the model on different fake content datasets compare to other transformer models?

RQ3: What can SHAP tell us about the behaviour of a model?

To fill the research gaps identified, a domain-adaptive, resource-efficient, and explainable misinformation detection model is suggested that is built on the DistilBERT architecture. The model is trained by the two-stage fine-tuning strategy. First, three common fake news datasets, namely, GossipCop, PolitiFact, and ISOT are trained. Afterward, domain adaptation is performed based on the BoolQ dataset, a series of binary factual question-answering examples. In this fashion, robustness of the model and its generalization capabilities can be assessed through the evaluation of the model across various textual domains.

A consolidated comparison of existing approaches across machine learning, deep learning, and transformer-based models, along with the proposed BiSplit-DistilBERT framework, is presented in Table 1. The table highlights datasets, performance, limitations, and research gaps, while also demonstrating the cross-domain evaluation capability of the proposed model using both fake news and factual question-answering data.

The proposed model is evaluated on three fake news datasets (PolitiFact, GossipCop, and ISOT) and one cross-domain factual question-answering dataset (BoolQ) to assess its generalization capability beyond misinformation detection tasks. As shown in Table 1 most existing approaches primarily focus on improving classification performance within specific domains, often lacking interpretability and cross-domain adaptability. In contrast, these limitations are addressed in the proposed BiSplit-DistilBERT model through the integration of a lightweight architecture, an adaptive early-exit mechanism, and a cross-domain evaluation strategy. A reduction in computational complexity is achieved while maintaining effective performance, making the approach more practical for real-time applications. Furthermore, model

Table 1 — Consolidated comparison of machine learning, deep learning, and transformer-based approaches for fake news detection with cross-domain evaluation

Model type	Algorithms used	Dataset(s)	Reported accuracy / performance	Key limitation	Unaddressed research gap
This work	BiSplit-DistilBERT (Proposed)	GossipCop, PolitiFact, ISOT, BoolQ	94.8%, 91.5%, 99%, 85.0%	Requires custom classification head; SHAP interpretability adds computation	None: simultaneously addresses efficiency, cross-domain generalization, and interpretability
ML-based	SVM[2]	Getting Real About Fake News	0.98	Dataset limited in size and diversity	Lack of contextual understanding and scalability
	Passive Aggressive, Naive Bayes, Logistic Regression, DT, LSTM, BERT[3]	Kaggle, GitHub (BERT News)	99.6% (DT), 99.8% Recall (LSTM)	High computational overhead for DL models	Absence of lightweight transformer comparisons; lacks scalability
	LR, SVM, Naive Bayes, RF, XGBoost, ANN[9]	Twitter, LIAR	Not reported	Focused only on basic ML models	Neglect of transformer-based architectures and generalization
	RF, MLP, DT, SVM[10]	UCI ML Repository	98.8% (MLP)	Poor real-world generalization	Dataset too limited for dynamic misinformation scenarios
	Content Similarity Measure (CSM)[11]	CoAID, FakeHealth, ReCOVery	91.06%	Focused on narrow domain	Ignores structural variation and explanation generation
DL-based	Custom CNN with Word Embedding[12]	COVID-19 Fake News	0.9619	Limited to COVID context	Lacks diversity and explainability
	GANM (Graph Attention with Memory[13]	Social Media Corpus	Not reported	Limited interpretability and scalability	No integration with transformers for attention mechanism
	Conv-FFD[14]	Chinese Rumor Corpus	Not reported	Domain-limited; hyperparameter tuning complexity	Domain-specific; lacks cross-language adaptability
	EfficientNet-BiLSTM[15]	YouTube-8M, Elsagate, NPDI Cartoon	Not reported	Not real-time; requires video segmentation	Not suitable for text-based misinformation detection
	CNNs, SVM, KNN[8]	Yonsei University CIP Lab (visual content)	89.5% (ResNet18)	Visual-only models; high GPU usage	Not applicable to text-based fake news; lacks multimodal fusion
	Domain-Adaptive Adversarial Training[7]	General news → Political/Financial	Not reported	Domain-specific only	No transparency; fails to generalize across diverse formats
Transformer-based	Dual BERT[16]	FakeNewsNet	0.854	Dataset size constraint	No domain-specific fine-tuning; lacks robustness
	BERT[17]	Twitter, Web Articles	Not reported	Early-stage prototype	No focus on explainability or low-resource applicability
	ARBERT, AraVec, CAMELBERT[18]	Arabic Fake News Datasets	Accuracy varies	High model loss; language-specific	Absence of universal architecture for multilingual misinformation
	RoBERTa-LightGBM[6]	Reuters.com fake news	95%	High computational cost	No lightweight design; lacks explainability

transparency is improved through the use of SHAP-based explanations, by which insights into prediction behaviour are provided. As a result, the model is considered suitable for deployment in real-world scenarios, particularly in resource-constrained settings and applications where reliable decision-making is critical.

Research Gap

Whereas much has been achieved in regards to fake content detection, most of the extant research has been confined to headline based or broad categorization of news. The contextual richness of contextual information, whether based on answering factual questions, opinion-based content

or tabloid style narratives, has been neglected. This has led to poor generalizability in the application of such models to other formats of misinformation. Moreover, model interpretability has not been given due attention, as a result of which, the process of decision making is frequently incomprehensible, which decreases the confidence in applications whose results are related to real life consequences. In solving these issues, a lightweight and interpretable model, known as BiSplit-DistilBERT, has been introduced in this paper. In contrast to traditional transformer models, which need a lot of computational resources, the given model is developed to work in resource-restricting settings. Two-stage fine-tuning approach has been employed wherein the initial training was performed on standard fake news datasets such as GossipCop, PolitiFact, and ISOT, and then it was adapted using the BoolQ dataset which included binary fact-based question-answer pairs. By doing so, better learning of finer semantic differences has been obtained leading to better cross-domain performance. Additional enhancements have been implemented in a binary splitting mechanism in the classification head. The prediction process has been partitioned into two phases rather than a single step classification and coarse-level categorization is then followed by a fined binary decision. This structure has made the classification task easier and has led to a faster convergence, lower cost of computation and better accuracy. SHAP-based explanations also have been used to improve model interpretability, where token-level contributions are visualized. The words that support the fake predictions are written in blue, whereas those that represent the true labels are written in orange, which enabled a better comprehension of the model choices. Good results have been noted in all benchmark datasets. Accuracy, precision, recall and F1-score values of 99% have been attained on ISOT. Accuracies of 94.8 and 91.5% have been reported on GossipCop and PolitiFact. With BoolQdataset, 85% accuracy, 84 F1-score, and 91 AUC have been achieved, which is better than BERT-base. These results show that an accurate, efficient and interpretable balance has been realized.

Methodology

In this research, a three-phase pipeline was used, including data preparation, model training, and evaluation. The architectural view of the proposed BiSplit-DistilBERT model and decision process is detailed in the relevant subsection, with the logical flow diagram depicting the decision process. The initial step was to develop a wide set of data by combining various publicly available databases, including the Fake and Real News dataset, GossipCop, and PolitiFact. A balanced distribution of fake and real news instances was obtained through this combination. Standard preprocessing steps were applied, and the complete dataset was tokenized using the DistilBERT tokenizer. The processed data was then divided into training and testing sets in an 80 to 20 ratio.

In the second phase, five transformer-based models, namely DistilBERT, BiSplit-DistilBERT, BERT-base, RoBERTa-base, and DeBERTa-base, were fine-tuned using the prepared dataset. Within BiSplit-DistilBERT, a hierarchical binary split classification mechanism was introduced, through which the decision process was simplified into smaller sub-tasks. To support this mechanism, a lightweight feed-forward classification head was integrated into the DistilBERT architecture. This component included a single hidden linear layer followed by a Rectified Linear Unit activation function, while no explicit activation was applied at the final output stage.

During forward propagation, the hidden representation was first transformed through a linear layer, after which non-linearity was introduced using the activation function. The output was then passed through another linear layer to obtain the final logits. The Cross Entropy Loss function was used during training, through which the softmax operation was handled internally. This structure guaranteed computational effectiveness and rendered the model applicable to be implemented in resource-restricted settings. In order to answer RQ1, DistilBERT is chosen as the base model because it has a lightweight and efficient model. DistilBERT has fewer parameters than BERT, but it still has a good contextual understanding based on knowledge distillation. This allows fewer computations and reduced memory usage and hence is appropriate in low resource environments. Moreover, it can find the semantic relationships in text, which makes it useful in detecting fake content, where subtle linguistic patterns and contextual information are of importance. The following features explain why DistilBERT was chosen to use in the proposed framework.

Architecture of the Proposed BiSplit-DistilBERT

The detailed architecture of the proposed BiSplit-DistilBERT framework, along with the adaptive early-exit mechanism, is illustrated in Fig. 1. The input text is tokenized with DistilBERT tokenizer, as a result of which a sequence of tokens is obtained, including the special CLS token and contextual word tokens. The tokens are then turned into contextual representations through the embedding layer, as the token representations, the segment representations, and the positional representations are merged. The encoded representations are then fed to the shared DistilBERT transformer encoder, which comprises of several encoder blocks with multi-head self-attention and position-wise feed-forward networks. Contextual relationships between tokens are captured with these layers and the rich features are represented. Adaptive inference is enabled in the proposed architecture by allowing a dynamic decision on whether further text processing is required. After contextual embeddings

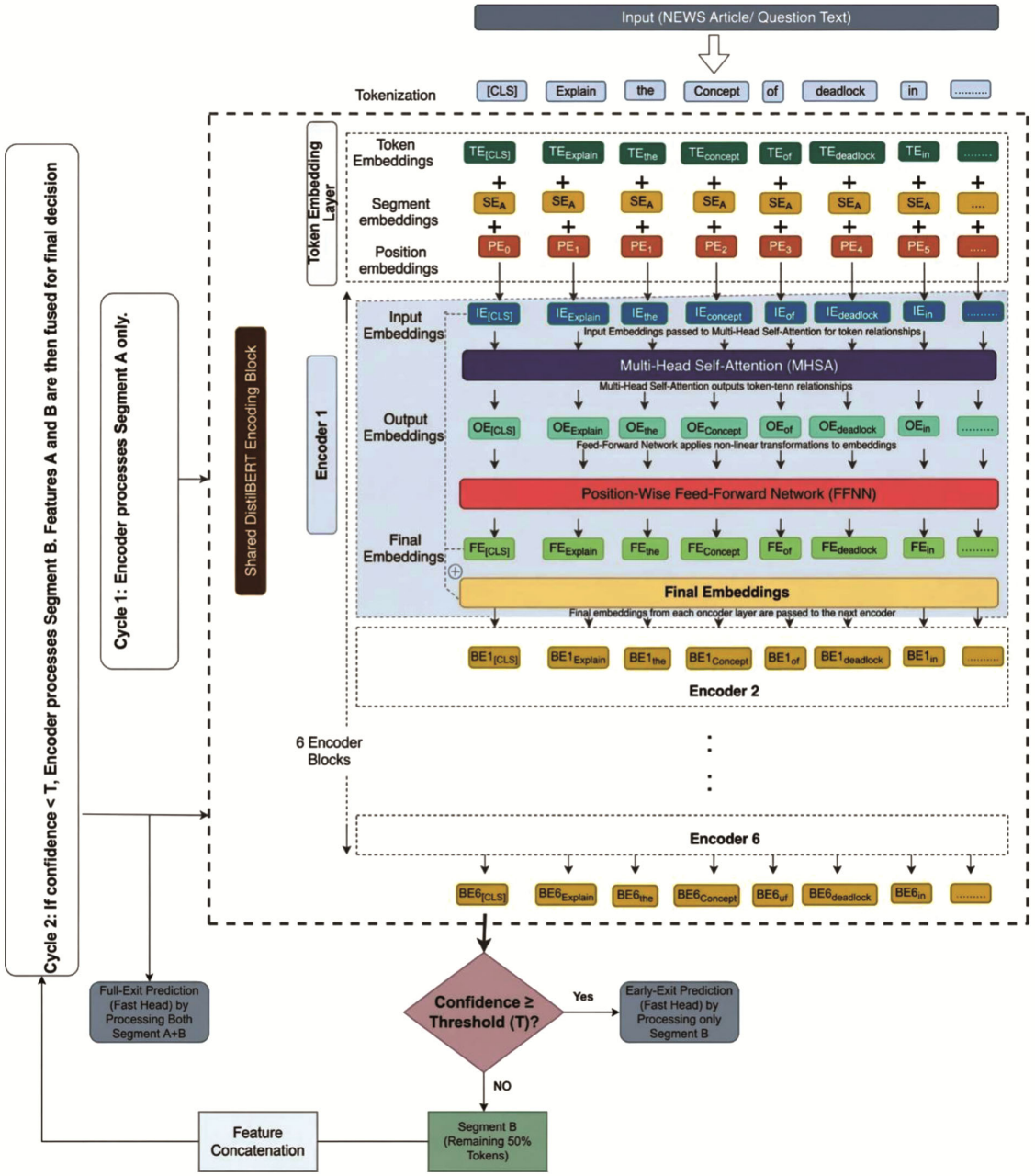

Fig. 1 — Token-level importance visualization using SHAP for an ISOT dataset sample

are obtained, the token sequence is divided into two parts. The first part, referred to as Segment A, contains the initial portion of the sequence, while the remaining tokens form Segment B. During the first stage of inference, only Segment A is processed, and a confidence score is computed using a lightweight classification head. If this confidence exceeds a predefined threshold, a prediction is generated at this stage, and further computation is avoided. In cases where the confidence is not sufficient, Segment B is also processed using the same encoder, and representations from both segments are combined to form a unified feature representation.

This combined output is then run through the final classification head, and a linear change and then a Rectified Linear Unit activation function transforms the result to give the final prediction. In this way, a decrease in computational effort is achieved without any impact on predictive performance.

This type of architecture is associated with a lower cost of computation both in training and inference and enhances faster convergence. Domain adaptation is conducted after the first training phase on the BoolQ dataset, a collection of yes or no questions that are based on facts. This further fine-tuning enhances the factual consistency of the model, and its generalization to other text formats. In the assessment of performance, the Accuracy, Precision, Recall, F1-score, and AUC are used to evaluate the performance. Moreover, memory usage, training time, and model size are also taken into account to determine the computational efficiency.

Performance Improvement through Adaptive Efficiency and Accuracy Enhancement

The BiSplit-DistilBERT model overall performance is enhanced by a decrease in effective computational workload, which is met by an adaptive early-exit inference mechanism. Simultaneously, the accuracy of classification is also increased with the help of the suggested design. The mechanisms that have been behind these improvements are explained in the subsequent sections.

Computational Efficiency

The proposed early-exit mechanism achieves this by cutting the effective workload of inference by processing a subset of the input samples after the initial segment (referred to as Segment A) is processed to decide whether to continue with further computation based on a predetermined confidence threshold τ. Where, p is the probability of the samples to meet the early-exit condition. The approximate computational cost is given as:

$$(1-p)\times 1 + p\times 0.5 = 1 - 0.5p \quad \ldots (1)$$

Here, p is the ratio between the number of samples that leave early and 50 is the half-computation, whereas 1 is the full-computation. It is noted that the total computation need reduces proportionally to the number of samples that drop out prematurely. As an illustration, when 40% of the inputs meet the early-exit condition, the expected cost of computation is 0.8, which translates to a processing effort decrease of about 20%. To elaborate more, take a scenario that has 10 input samples. Assuming that 6 samples are subject to complete processing and 4 samples are subject to partial processing, the cumulative processing of 8 complete inputs is equal to the workload of processing the 6 samples that are completely processed. By doing so, even in the case of early exit being applied to a subsets of the data, computational saving is obtained. Because no further dataset as benchmarking of complete runtime has been performed in this revision, the efficiency improvement is reported as a measure of estimated computation savings instead of direct latency measurements.

Accuracy Enhancement

Along with the use of computational efficiency, the BiSplit-DistilBERT model also includes classification accuracy improvement due to the structured and confidence-conscious decision-making process. In comparison to the traditional transformer models, when the classification is carried out by the classification in one space, hierarchical coarse-to-fine strategy is considered in the proposed approach. By using binary-split classification head, the decision space is reduced to smaller intermediary tasks, not only enhancing the overall separability of classes, but also enabling subtle semantic variations to be better represented. One of the main contributions to this enhancement is made by the confidence-based control mechanism, which prevents premature or wrong predictions. In case of a lack of the confidence provided by the first part, which is called Segment A, the decision is not made, and the rest part, Segment B, is run through. The obtained segment representations are then merged to create a more enriching and informative feature space. In this adaptive combination, superior forecasts are undertaken particularly when the input is uncertain or intricate. To illustrate, in a sentence like Doctors say turmeric milk

may help improve immunity, the first part might include words that might drive a prejudiced forecast to an authoritative label. In the proposed model, however, the deficit of confidence is identified at this stage and the whole text is reviewed prior to the final decision being taken, thereby reducing risk of misclassifying the text. Additionally, high confidence cases can be exited at an early stage, without the needless processing of less informative tokens in further stages. The approach allows a moderate process of decision making because it incorporates early signal detection with more contextual information into the system leading to increased robustness and predictive accuracy. This can be attributed to experimental results on the BoolQ dataset, where an accuracy of 85% was achieved as opposed to 78% of the BERT-base model.

Dataset used

A detailed description of the datasets used in this study shown in Table 2. Fake news detection is done with three datasets (PolitiFact, GossipCop, and ISOT), and the BoolQ dataset is used as cross-domain factual question-answer benchmark to assess whether the proposed model can be generalized out of misinformation detection tasks.

Dataset Distribution

To represent the distribution of tokenized text lengths for fake and true labels across four benchmark datasets, namely BoolQ, GossipCop, ISOT, and PolitiFact, KDE plotsare used , shown in Fig. 2. A consistent pattern has been observed, in which shorter text lengths are generally associated with fake news instances, while longer text sequences are more commonly linked to true news across all datasets. This difference in length distribution is reported as an exploratory observation at the dataset level.

It should be noted that text or article length has not been explicitly used as a handcrafted feature in the proposed BiSplit-DistilBERT model. Instead, training has been performed directly on tokenized text sequences, allowing meaningful contextual patterns to be learned from the content itself. Further, smoothed KDE curves have been used instead of the conventional histograms because better visual clearness has been attained thus more appropriate in the analysis of finer details and publication-ready analysis. The general descriptions of the datasets of PolitiFact, GossipCop, ISOT and BoolQ do not explicitly mention that fake news is shorter than real news. The difference in length observed is not an intrinsic dataset definition, but an empirical phenomenon that we observe in our analysis.

This composite in Fig. 2 illustrates the distribution of title lengths for fake and real news articles across four datasets: GossipCop, Politifact, ISOT and BoolQ Dataset

Logical Flow Diagram

The essence of decision-making involved in the BiSplit-DistilBERT framework of fake content detection can be seen in Fig. 3, In this work, every input text instance is first split into two parts. The initial half is coded and assessed by a conditional check of confidence. In case enough confidence is reached, the prediction is obtained through the first half of the text only. Where the confidence level is insufficient both halves are coded, merged and utilized to create a joint feature representation. The

Table 2 — Datasets used in the study

Dataset name	Task	Classes	Number of total entries	Attributes	Number of true entries	Number of fake entries	Description
PolitiFact	Fake News Detection	True / Fake	622	id, news url, title, tweet ids	314	308	From PolitiFact, a fact-checking website for claims made by public figures and politicians.
GossipCop	Fake News Detection	True / Fake	5,000	id, news url, title, tweet ids	2,500	2,500	From GossipCop, dedicated to verifying celebrity and entertainment news.
ISOT (Information Security and Object Technology)	Fake News Detection	True / Fake	44,898	title, text	21,417	23,481	A benchmark dataset containing real and fake news articles used widely for misinformation detection.
BoolQ	Question Answering (Cross-Domain Evaluation)	Yes / No	11,140	text, label	6,809	4,331	From BoolQ, a dataset of yes/no factual questions drawn from real-world sources used to evaluate cross-domain generalization.

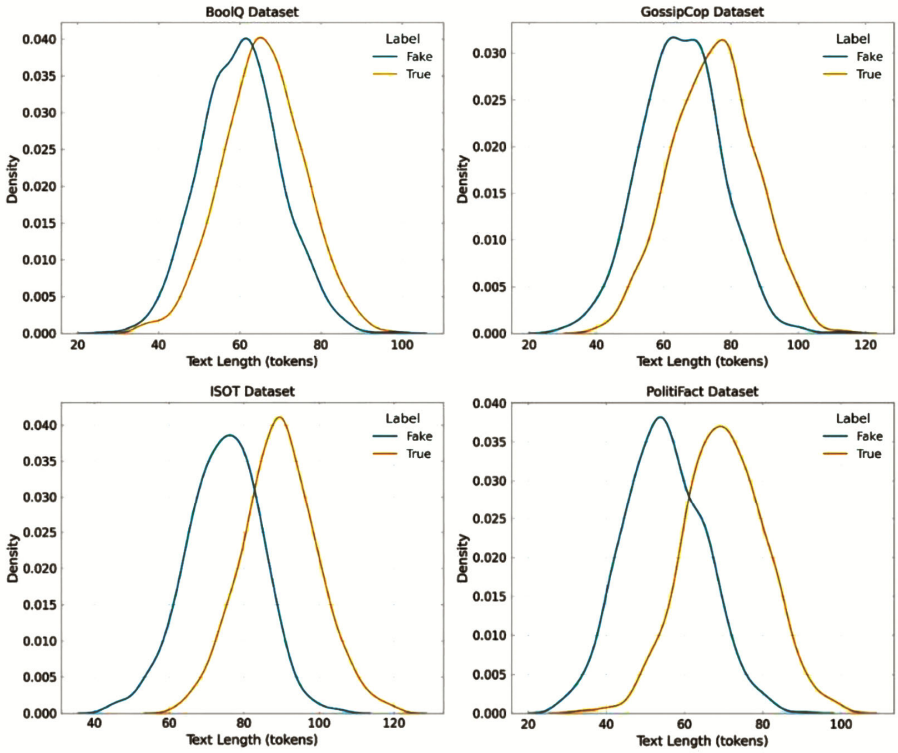

Fig. 2 — Evaluation metrics for DistilBERT model: ROC curve and confusion matrix

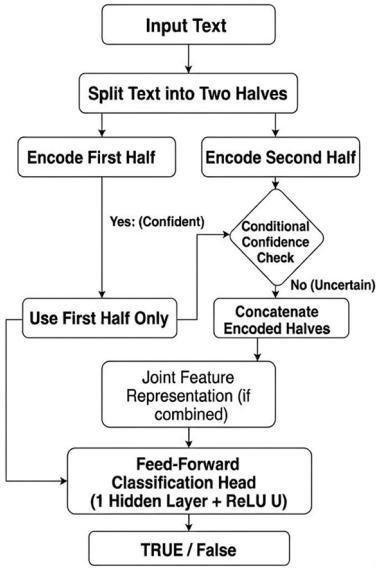

Fig. 3 — Evaluation of BiSplit-DistilBERT against transformer models on fake news datasets

resulting combined representation is then fed through a simple feed-forward classification head made up of a single hidden layer with a Rectified Linear Unit activation function to generate the final prediction.

An early-exit strategy is implemented, where only first part of the input text is handled at the first level. The input is categorized when there are high pointers of fake content in this segment without any additional analysis. This observation is in line with the support this behaviour has because fake news tends to have deceptive cues during the initial part of the text, but in the cases when the first part fails to give enough evidence, the rest of the text is also examined to make sure that the classification will prove to be more accurate.

This adaptive inference strategy minimizes computation and does not affect the accuracy of prediction. This decision process is preceded by the entire pipeline, such as preprocessing, tokenizing with the DistilBERT tokenizer, training on mixed fake news datasets, and then fine-tuning with factual data such as the BoolQ dataset. It is assessed using Accuracy, Precision, Recall, F1-score and AUC, and overall, we can notice some regular improvement when compared to the baseline models, like DistilBERT and DeBERTa.

Algorithm: Fake Content Detection using BiSplit-DistilBERT Overall flow

Load Data:

True articles $Dt=\{(x_i,1)\}$

Fake articles $Df=\{(x_i,0)\}$

Data: News articles $D=\{(x_i,y_i)\}$, where $y_i= 1$ for true, $y_i= 0$ for fake

Combine Datasets:

$D=D_t \cup D_f$

Split Data:
Train set D_{train} (80%)
Test set D_{test} (20%)
Prepare Data:
Tokenize articles using DistilBERT tokenizer: T(x)
Create Custom Datasets:
Class NewsDataset:
Store tokenized articles T(x) and labels y
idx: Return $T(x_{idx}),y_{idx}$
length: Return |y|
Training dataset D_{train}
Test dataset D_{test}
Set Up Model and Training Settings:
Load pre-trained BiSplit-DistilBERT model for classification (modified classification head based on DistilBERT)
Training settings: Save path, epochs=3, batch_size=16, warmup steps, weight decay, logging
Train Model:
Trainer with model, settings, Dtrain, Dtest
Train model
Fine-Tune Model:
Fine-tune the trained BiSplit-DistilBERT model using domain-specific factual QA data (BoolQ)
Make Predictions:
Use of trained model on Dtest
Outputs convert to predicted labels $\hat{y}$
Evaluate Model:
Evaluation metrics calculate including accuracy, precision, recall, F1-score, and
AUC.

To improve clarity and reproducibility, the implementation of the proposed BiSplit early-exit mechanism is formally described in below Algorithm

Algorithm: BiSplit Early-Exit Mechanism

Input: Tokenized input sequence T(x), confidence threshold τ
Output: Predicted label y

1. Split T(x) into two segments:
Segment A ← first 50% tokens
Segment B ← remaining tokens

2. Encode Segment A using DistilBERT encoder
3. Compute prediction probabilities P_A

4. If max(P_A) ≥ τ then
Return predicted label (Early Exit)
5. Else
Encode Segment B using shared encoder
Concatenate representations of Segment A and Segment B
Pass through final classification head
Return final predicted label

Above algorithm outlines the step-by-step execution of the proposed mechanism, demonstrating how partial input processing enables efficient inference while maintaining classification accuracy.

A simplified conceptual implementation of the proposed BiSplit early-exit mechanism (Algorithm 2) is presented below.

```
defbisplit_forward(input_tokens, model, threshold=0.8):
defbisplit_forward(input_tokens, model, classifier_head, threshold=0.8):
 # Step 1: Split input into two segments
split_idx = len(input_tokens) // 2
segment_A = input_tokens[:split_idx]
segment_B = input_tokens[split_idx:]
 # Step 2: Encode Segment A
feat_A = model(segment_A)
probs_A = softmax(feat_A)
 # Step 3: Early-exit condition
 if max(probs_A) >= threshold:
 return argmax(probs_A)
 # Step 4: Encode Segment B
feat_B = model(segment_B)
 # Step 5: Combine representations
combined_features = concatenate(feat_A, feat_B)
 # Step 6: Final prediction
final_logits = classifier_head(combined_features)
final_probs = softmax(final_logits)
returnargmax(final_probs)
```

In Fig. 4, a comparative visualization of word clouds of common words observed in fake and real news articles is depicted. Emotional and sensational words like BREAKING, Trump, Obama and Killed stand out in the word cloud that is associated with fake news. Conversely, the word cloud that reflects real news is marked by the use of more formal and structured language, and the words Transcript, Remarks, President, Congress are more common. By this visual analogy, it is evident how the linguistic style and tone differ between the deceptive and non-deceptive reporting in online media. It is noted that fake news is more prone to the use of attention-grabbing, emotionally-charged expressions, and real news is linked to the use of a more formal and informative language. These observations offer valuable information on the difference in presentation of content in various kinds of news articles.

Explainable AI (XAI) Integration with BiSplit-DistilBERT

To enhance transparency and interpretability of the proposed BiSplit-DistilBERT framework, SHAP was integrated as a post-hoc explainability technique. SHAP is a game-theoretic approach that assigns an importance score to each input feature by estimating its marginal contribution to the model prediction.

The output of the explainable AI component, in which SHAP values are used to interpret predictions for fake and real news classification, is presented in Fig. 5. In the input text, the tokens are highlighted based on their contribution to the final prediction. It is

Fig. 4 — SHAP visualization of token contributions in the proposed BiSplit-DistilBERT model

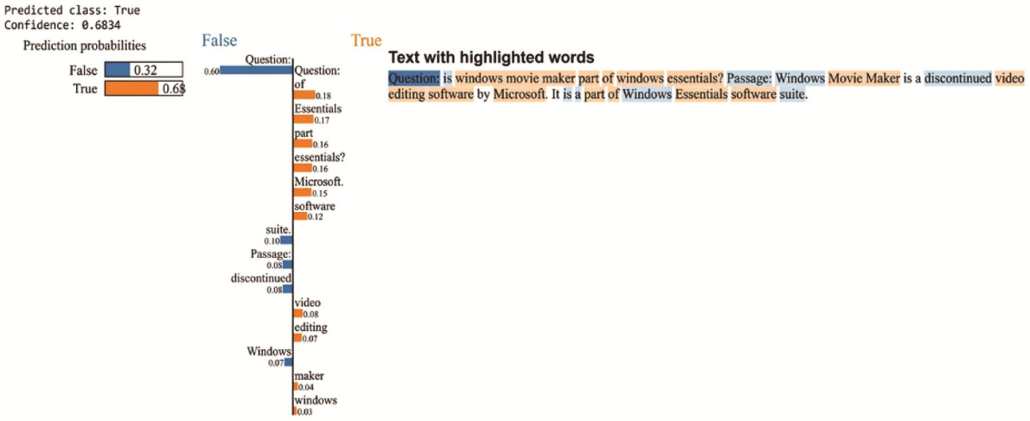

Fig. 5 — Word cloud comparison between fake and real news

noted that the tokens like shares, youngest and Christmas have positive contribution and shift the prediction to the true class but the tokens like rare and photo have negative contribution and shift the prediction to the false class. In the case of the example sentence, Paris Jackson shares rare photo of her youngest brother Blanket on Christmas, the estimated probability of the True and Fake classes are 0.52 and 0.48 respectively. The base value represents the expected model output in the absence of input features. Through this visualization, improved transparency and interpretability of the decision-making process are achieved.

In this study, SHAP is applied after the training phase to interpret predictions generated by the fine-tuned BiSplit-DistilBERT model. Contextual embeddings for each token are first produced by the DistilBERT encoder. These embeddings are then processed by the binary-split classification head to obtain class probabilities for fake and true categories.

SHAP values are calculated based on the prediction probabilities of the trained model to analyses the contribution of individual tokens. A text model based on transformers is estimated to predict the effect of every token with the help of an explainer. Background dataset is used to predict a baseline value of prediction, which is the expected output without the presence of a given input token. The trained model is fed tokenized input during the explanation process and prediction probabilities are given. The contribution of individual tokens is then approximated by assessing the effects of masking or perturbing model output, relative to the baseline. The tokens, which have the positive SHAP value increase the likelihood of the predictive class and those with the negative value reduce the prediction confidence. SHAP does not change the training process or enhance the accuracy directly, but the identification of influential tokens can give valuable insights into model behavior. Through this interpretability layer, the meaningful textual patterns are learned and are therefore increased transparency and trust to the presented framework.

Results and Discussion

This section contains, the comparative analysis is provided to analyse the performance of the BiSplit-DistilBERT model in comparison with the base transformers models in terms of not only the classification but also in the computational efficiency. It

is evaluated using four benchmark data, i.e., GossipCop, PolitiFact, ISOT and BoolQ. The standard performance measures, such as Accuracy, Precision, Recall, F1-score and AUC, are implemented to guarantee a moderate evaluation of both accuracy and robustness in classification. As shown in Fig. 6 the highest performance under all datasets is done by the BiSplit-DistilBERT model. It achieves an accuracy of 85% on the BoolQ dataset and an accuracy of 94.8%, 91.5% and 99% on GossipCop, PolitiFact and ISOT respectively. These findings reveal consistent results compared with baseline models and show the effectiveness of the proposed approach on datasets with different features.

Evaluation Metrics

In order to measure the classification performance of the transformer models, standard measures were adopted. Recall was taken into account to determine how effective the model was in identifying the true positives correctly and at all the difficulty levels simple, average, and tough. It was calculated using the following formula:

$$Recall = \frac{TP}{TP+FN} \quad \ldots (2)$$

where, TP represents the true positives and 2x 3 represents the false negatives. The F1-score was applied to give a balanced value of precision and recall of the model particularly where there may be uneven distributions of classes.

It is the harmonic average of precision and recall and it is described as:

$$\text{F1-score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad \ldots (3)$$

where, Precision is calculated as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

TP in this case refers to true positives and FP refers to false positives. An increased F1-score means that the model has a more favorable balance between precision and recall, meaning that it is more likely to accurately predict the positive instances and reduce the number of false predictions.

The overall correctness of the predictions of the model was determined through accuracy. It was defined as:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \ldots (4)$$

Here, TN and FP are true negatives and false positives, respectively. AUC was used to quantify the model capability to differentiate between fake and real news.

The larger the AUC value, the greater was the discrimination ability.

The AUC was computed as:

$$AUC = \int_0^1 TPR \, dFPR \quad \ldots (5)$$

In this context, TPR (true positive rate) and FPR (false positive rate) were utilized to construct the ROC curve.

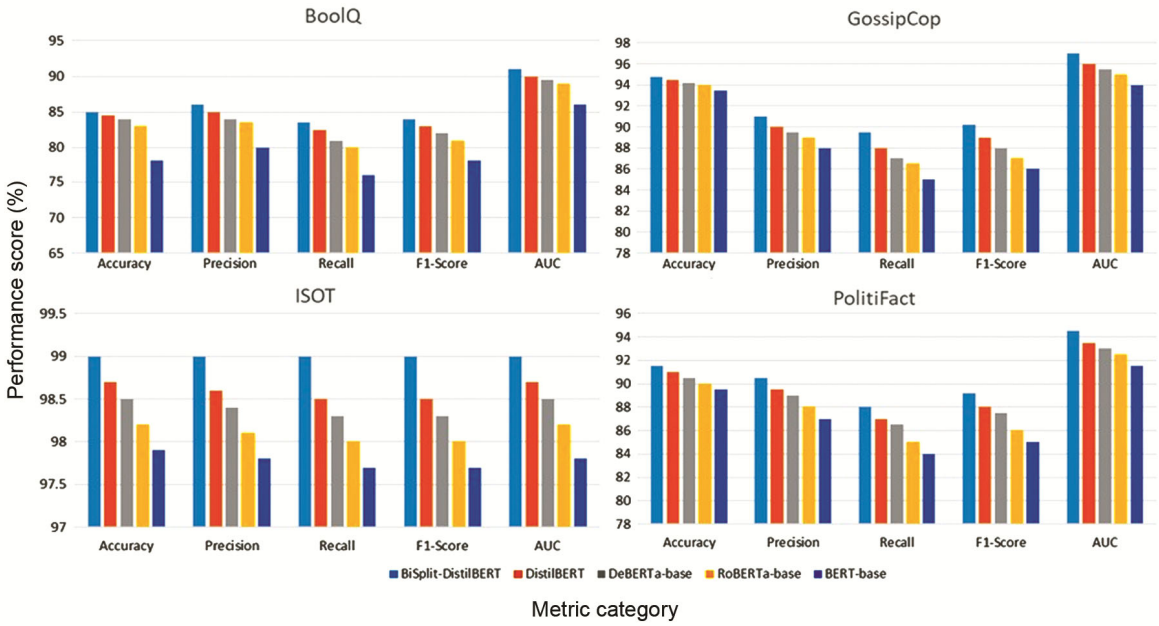

Fig. 6 — Logical flow of the proposed BiSplit-DistilBERT architecture

Comparative Evaluation of the DistilBERT, DeBERTa, and RoBERTa Models

A detailed comparison between the proposed BiSplit-DistilBERT model and baseline transformer models, including DistilBERT, DeBERTa-base, RoBERTa-base, and BERT-base, is provided in Fig. 6 and Table 3, on four benchmark datasets, namely BoolQ, GossipCop, ISOT, and PolitiFact. Performance is assessed using standard evaluation measures like Accuracy, Precision, Recall, F1-score, and AUC. Based on the obtained results it can be observed that consistent gains are obtained with the BiSplit-DistilBERT model on all datasets. The results on the ISOT dataset are nearly flawless, with scores of 99% in all measurements. Likewise, a greater accuracy of 94.8% and 91.5% are achieved on the GossipCop and PolitiFact datasets, respectively, and better precision and F1-scores are found compared to baseline models. On the BoolQ dataset, which is more complex, an accuracy of 85% and an AUC of 91 can be reached, showing definite progress of the BERT-base model. The figure gives a visual comparison of the model performance, whereas the table gives precise numbers, thus enhancing clarity and making redundancy in interpretation redundant.

Besides that, the effectiveness of the proposed binary-split mechanism is compared to that of DistilBERT and underscores its efficiency. Early predictions are made by splitting the input into pieces before using all the information necessary, with a more in-depth analysis carried out only where needed. This is an adaptive mechanism, which minimizes the computational effort with no predictive cost. The interpretability further improves with SHAP in which token-level contributions are plotted. Tokens that support the fake predictions are indicated in blue whereas tokens that support true predictions are indicated in orange that gives a better understanding of the behavior of the model. It is also observed that cross-domain generalization is tested on the BoolQ dataset. The good performance seen on this dataset implies that robustness is attained in a variety of textual situations.

Two important evaluation tools that are shown in Fig. 7. employed to evaluate the performance of the BiSplit-DistilBERT model in misinformation detection. The ROC curve above on the left gives an AUC value of 0.99, which translates into a high ability of the model to separate fake and real news cases. The confusion matrix heatmap on the right side demonstrates the detailed results of the classification, 4,950 true positives and 4,950 true negatives are correctly recognized, only 50 false positives and 50 false negatives are found. This is due to the two-stage fine-tuning approach that is embraced in the model. Within this technique, the model is initially trained on the tasks of classifying fake news and then further trained on specific domain factual question-answering data. In this way, better generalization and understanding the context are attained. Besides that, the internal restructuring of the classification layer

Table 3— Performance comparison and ablation analysis of BiSplit-DistilBERT across multiple datasets

Dataset	Model	Accuracy	Precision	Recall	F1-Score	AUC
	BiSplit-DistilBERT	85	86	83.5	84	91
	DistilBERT	84.5	85	82.5	83	90
BoolQ (Cross Domain)	DeBERTa-base	84	84	81	82	89.5
	RoBERTa-base	83	83.5	80	81	89
	BERT-base	78	80	76	78	86
	BiSplit-DistilBERT	94.8	91	89.5	90.2	97
	DistilBERT	94.5	90	88	89	96
GossipCop	DeBERTa-base	94.2	89.5	87	88	95.5
	RoBERTa-base	94	89	86.5	87	95
	BERT-base	93.5	88	85	86	94
	BiSplit-DistilBERT	99	99	99	99	99
	DistilBERT	98.7	98.6	98.5	98.5	98.7
ISOT	DeBERTa-base	98.5	98.4	98.3	98.3	98.5
	RoBERTa-base	98.2	98.1	98	98	98.2
	BERT-base	97.9	97.8	97.7	97.7	97.8
	BiSplit-DistilBERT	91.5	90.5	88	89.2	94.5
	DistilBERT	91	89.5	87	88	93.5
PolitiFact	DeBERTa-base	90.5	89	86.5	87.5	93
	RoBERTa-base	90	88	85	86	92.5
	BERT-base	89.5	87	84	85	91.5

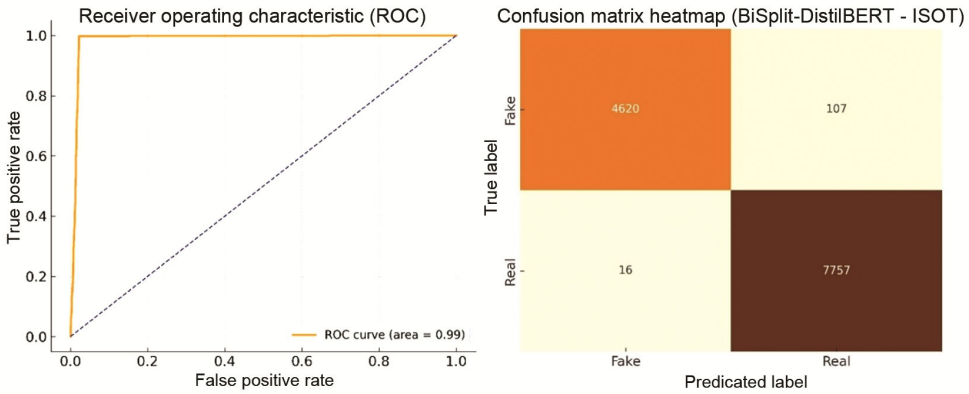

Fig. 7—Histogram of title/text lengths in fake and real samples across datasets

Table 4 — Representative examples of model predictions across datasets

Dataset	Input Text Snippet	Ground Truth	Model Prediction	Confidence / Early Exit?
ISOT	"BREAKING: Putin Directly Orchestrated Hacks To Get Trump Elected (VIDEO)"	Fake	Fake	98.2% / Yes
PolitiFact	"Obama administration opens the door for states to seek major changes in welfare-to-work law"	Real	Real	94.5% / No
GossipCop	"Prince Harry Takes Meghan Markle for Tea at the Palace to Meet Queen Elizabeth: Report"	Real	Real	91.8% / No
BoolQ	"Has anyone won the grand slam in golf in one year? True."	TRUE	TRUE	97.6% / Yes

facilitates the improvement in the performance of the classification, as better separation of the classes is ensured. Consequently, the boundaries of decisions are reinforced, which brings about increased strength of the model. These findings, in general, indicate a high level of reliability of the proposed solution, as well as a balanced trade-off between precision and recall. The results also show that the model can be successfully implemented in practice in the context of detection of fake news in the real world.

Qualitative Analysis and Model Interpretability

As good performance is reflected in the quantitative measures, the qualitative analysis is necessary to comprehend the decision-making procedure of the BiSplit-DistilBERT model and the effectiveness of the early-exit mechanism in practice. To this end, the samples of each of the four benchmark datasets are chosen randomly to study the treatment of various textual structures and linguistic patterns. This analysis leads to getting an idea of the correlation between the textual content and the confidence of the model predictions and the early-exit pathway activating. This analysis gives a better insight on how the model can modify its inference mechanism depending on the type of input information.

Representative samples from the four datasets used in this study, namely ISOT, PolitiFact, GossipCop, and BoolQ, are presented in Table 4, to illustrate how this model classifies. The ground truth is the original and verified label of each news item or statement, where a label of Fake denotes a deceptive one and label Real or True denotes a verified one. The model prediction is the category of the BiSplit-DistilBERT architecture. The last column which is known as Confidence or Early Exit is an indicator of efficiency of decision-making process. The percentage value is the softmax probability which is the degree of confidence of the prediction. The indication of 'Yes' or 'No' specifies whether the early-exit mechanism is activated. In case of Yes, the confidence threshold is predefined and only the initial part of the input text is used to reach the threshold. When the indicator No is noted, the entire text is handled and a faithful forecast is generated.

Presented in Fig. 8, through which an explainable AI interpretation of the model's behavior on a sensational headline is provided. A heatmap is generated using SHAP, where importance is assigned to individual tokens based on their contribution to the final prediction.

It is observed that tokens highlighted in red, such as "BREAKING" and "Putin," contribute strongly toward a 'Fake' classification. In contrast, tokens that support a 'Real' classification are represented in blue

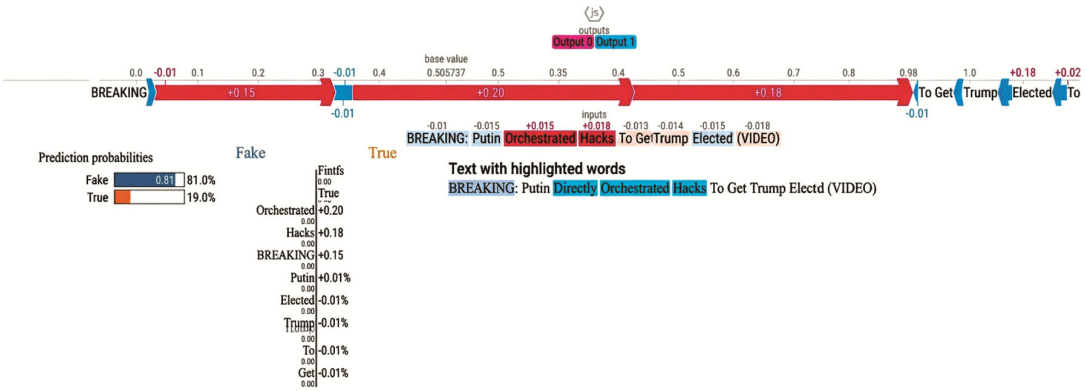

Fig. 8 — Token-level importance visualization using SHAP for an ISOT dataset sample

Table 5 — Benchmarking BiSplit-DistilBERTwith existing misinformation detection approaches

Model used	Dataset(s)	Accuracy (%)	Key limitations
BiSplit-DistilBERT + SHAP + BoolQ (This Work)	ISOT, PolitiFact, GossipCop, BoolQ	99 (ISOT), 94.8 (GossipCop), 91.5 (PolitiFact), 85 (BoolQ)	Lightweight architecture was utilized; interpretability was enhanced through SHAP; improved AUC and F1-scores were recorded across multiple datasets.
Dual BERT with MaxWorth span selection[16]	FakeNewsNet	85.4	Cross-domain evaluation was not performed; reliance was placed on ClaimBuster API and headline-body structure.
Blockchain + Smart Contracts for Smart Grid[19]	Simulated Smart Grid (20 residents)	N/A	A real-world dataset was not used; NLP applicability was excluded; only sequential energy logic was implemented.
TriFN (tri-relational model with content-user-publisher)[20]	FakeNewsNet (BuzzFeed&PolitiFact)	86.4 (BuzzFeed), 87.8 (PolitiFact)	Metadata on user credibility and publisher bias was required; interpretability was restricted due to matrix factorization.
CSI (RNN-based hybrid model using user behavior)[21]	Twitter, Weibo	89.2 (Twitter), 95.3 (Weibo)	Model decisions were influenced by user behavior and propagation patterns; token-level interpretability was not incorporated; generalization across domains was limited.

when present. A clear concentration of these high-impact tokens is identified at the beginning of the text.

This distribution of influential features supports the design of the early-exit mechanism in the BiSplit-DistilBERT model. Since the most significant linguistic cues are captured within the initial portion of the input, a confident prediction can be generated at an early stage. As a result, further processing is avoided, leading to reduced computational effort without affecting predictive accuracy.

Comparison with Existing Work

In order to contextualize the performance of the suggested BiSplit-DistilBERT model, its performance is compared with a set of state-of-the-art fake news detection methods, which have been reported in recent literature. These models are evaluated according to their reported accuracy, datasets employed, and their limitations as to architecture or operations. As shown in Table 5, he proposed framework leads to significant gains in predictive performance, whereas interpretability and low-computational costs are preserved, which are not sufficiently taken care of in previous methods. Despite competitive outcomes that are presented by the current transformer-based approaches, some limitations are found. A dual BERT architecture using headline and body pairs, which provides an accuracy of 85.4%, does not include domain adaptation and multi-task learning techniques and so is restricted in its generalization ability. The other model is reported to have an accuracy of 88% but depends heavily on the user behaviour and the patterns of content propagation making it less dependent on the linguistic attributes. Besides these, there are also more sophisticated methods that claim over 93% accuracy, but do not provide token-level interpretability and performance-efficient deployment, limiting their

usage in real-time contexts. Conversely the proposed model is initially trained on several fake news datasets, such as ISOT, PolitiFact and GossipCop and then fine-tuned on the BoolQ dataset, which consists of factual question-answer data. On ISOT, GossipCop, PolitiFact, and BoolQ, 99%, 94.8%, 91.5%, and 85% accuracies are obtained, respectively. On ISOT an AUC of 99 is observed, with values of 97, 94.5 and 91 observed on the other datasets, suggesting a high level of discriminative ability. Although fewer parameters are used than full-sized BERT models, they are still able to perform competitively. Class imbalance can be improved with the help of changes in the classification layer, and token-level interpretability can be assisted with SHAP. On balance, the comparative analysis demonstrates that the proposed BiSplit-DistilBERT model performs the best in accuracy and AUC, as well as, it tackles key issues like domain adaptability, interpretability, and efficient deployment, which makes it applicable to real-time tasks of misinformation detection.

Conclusions

This paper proves that BiSplit-DistilBERT is an appropriate model to respond to the necessity to use fewer resources and detect misinformation that can be interpreted. The framework combines a binary-split classification head with a flexible early-exit mechanism to balance high predictive accuracy and less computational workload, and it can be deployed in resource-constrained settings. The cross-domain test on the factual question-answering data also supports the fact that the model is able to generalize outside of the typical news articles. Also, the inclusion of SHAP-based explanations promotes transparency and makes sound decision-making. The study has a limitation because it uses curated benchmark datasets, which might not be completely indicative of misinformation changes in real-world scenarios. The future research could look at how the streaming data can be combined and how the framework can be applied in specialized fields like healthcare and legal use. This paper is a step in the right direction of creating a scalable and reliable method of detecting misinformation.

References

1 Jamshidi S, Mohammadi M, Bagheri S, Najafabadi H E, Rezvanian A, Gheisari M, Ghaderzadeh M, Shahabi A S & Wu Z, Effective text classification using bert, mtmlstm and dt, *Data Knowl Eng*, 151 (2024) 102306, doi:10.1016/j.datak.2024.102306.

2 Baarir N F & Djeffal A, Fake news detection using machine learning, *Proc 2nd Int Work Human-Centric Smart Environ Heal Well-Being (IHSH)*, (2021) 125–130, doi:10.1109/IHSH51661.2021.9378748.

3 Pal A, Pranav & Pradhan M, Survey of fake news detection using machine intelligence approach, *Data Knowl Eng*, 144 (2023) 102118, doi:10.1016/j.datak.2022.102118.

4 Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z & Fu J, Pre-trained language models in biomedical domain: A systematic survey, *ACM Comput Surv*, 56 (2023) Article 165, doi:10.1145/3611651.

5 Junaid T, Sumathi D, Sasikumar A N, Suthir S, Manikandan J, Khilar R, Kuppusamy P G & Janardhana Raju M, A comparative analysis of transformer-based models for figurative language classification, *Comput Electr Eng*, 101 (2022) 108051, doi:10.1016/j.compeleceng.2022.108051.

6 Srinivasan D, Roberta-lightgbm: a hybrid model of deep fake detection with pre-trained and binary classification, *Infocomp*, 23(1) (2024) 1–11, https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/3060.

7 Ng K C, Ke P F, So M K P & Tam K Y, Augmenting fake content detection in online platforms: a domain adaptive transfer learning via adversarial training approach, *Prod Oper Manag*, 32 (2023) 2101–2122, doi:10.1111/poms.13959.

8 Noah A S, Ghannam N E, Elsharawy G A & Desuky A S, An intelligent system for detecting fake materials on the internet, *Int J Mod Educ Comput Sci*, 15 (2023) 42–59, doi:10.5815/ijmecs.2023.05.04.

9 Alghamdi J, Luo S & Lin Y, A comprehensive survey on machine learning approaches for fake news detection, *Multimed Tools Appl*, 83 (2024) 51009–51067, doi:10.1007/s11042-023-17470-8.

10 Yousaf K & Nawaz T, A deep learning-based approach for inappropriate content detection and classification of youtube videos, *IEEE Access*, 10 (2022) 16283–16298, doi:10.1109/ACCESS.2022.3147519.

11 Buchner J, Playing an augmented reality escape game promotes learning about fake news, *Technol Knowl Learn*, 30 (2025) 425–445, doi:10.1007/s10758-024-09749-y.

12 Akhter M, Hossain S M M, Nigar R S, Paul S, Kamal K M A, Sen A & Sarker I H, Covid-19 fake news detection using deep learning model, *Ann Data Sci*, 11(6) (2024) 2167–2198, doi:10.1007/s40745-023-00507-y.

13 Chang Q, Li X & Duan Z, Graph global attention network with memory: A deep learning approach for fake news detection, *Neural Netw*, 172 (2024) 106115, doi:10.1016/j.neunet.2024.106115.

14 Zhang Q, Guo Z, Zhu Y, Vijayakumar P, Castiglione A & Gupta B B, A deep learning-based fast fake news detection model for cyber-physical social services, *Pattern Recognit Lett*, 168 (2023) 31–38, doi:10.1016/j.patrec.2023.02.026.

15 Blankenship R J, Deep fakes, fake news, and misinformation in online teaching and learning technologies, *IGI Global*, (2021).

16 Farokhian M, Rafe V & Veisi H, Fake news detection using dual bert deep neural networks, *Multimed Tools Appl*, 83 (2024) 43831–43848, doi:10.1007/s11042-023-17115-w.

17 Al Ghamdi M A, Bhatti M S, Saeed A, Gillani Z & Almotiri S H, A fusion of bert, machine learning and manual

approach for fake news detection, *Multimed Tools Appl*, 83 (2024) 30095–30112, doi:10.1007/s11042-023-16669-z.

18 Azzeh M, Qusef A & Alabboushi O, Arabic fake news detection in social media context using word embeddings and pre-trained transformers, *Arab J Sci Eng*, 50 (2025) 923–936, doi:10.1007/s13369-024-08959-x.

19 Khattak H A, Tehreem K, Almogren A, Ameer Z, Din I U & Adnan M, Dynamic pricing in industrial internet of things: blockchain application for energy management in smart cities, *J Inf Secur Appl*, 55 (2020) 102615, doi:10.1016/j.jisa.2020.102615.

20 Shu K, Wang S & Liu H, Beyond news contents: the role of social context for fake news detection, *Proc ACM Int Conf Web Search Data Min*, (2019) 312–320, doi:10.1145/3289600.3290994.

21 Ruchansky N, Seo S & Liu Y, CSI: A hybrid deep model for fake news detection, *Proc ACM Int Conf Inf Knowl Manag*, (2017) 797–806, doi:10.1145/3132847.3132877.