JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Cross-Domain Faithfulness Evaluation of SHAP and Attention-Based Explanations in Transformer NLP Models

Dony Bahtera Firmawan * and Brian Rizqi Paradisiaca Darnoto

Informatics Department, University of Jember, Jember 68121, Indonesia;
e-mail : donybf@unej.ac.id; brianrizqi@unej.ac.id
* Corresponding Author : Dony Bahtera Firmawan

Abstract: Transformer-based models such as BERT, RoBERTa, DistilBERT, and DeBERTa have achieved remarkable performance across a wide range of natural language processing (NLP) tasks. However, their decision-making processes remain difficult to interpret, particularly in high-risk applications such as hate speech detection, where unreliable explanations may undermine model transparency, trust, and accountability. This study investigates whether explainability methods remain faithful and stable under domain shift in transformer-based text classification. Four transformer architectures were fine-tuned and evaluated on two linguistically distinct datasets: IMDb Movie Reviews and Hate Speech Offensive. Model performance and explanation quality were assessed using classification accuracy, macro F1-score, top-k token-removal faithfulness analysis, and cross-domain Spearman rank correlation. Experimental results show that DeBERTa achieved the highest classification performance, reaching accuracies of 95.6% on IMDb and 91.3% on Hate Speech. Across all evaluated models and datasets, SHAP consistently produced higher faithfulness scores than attention-based explanations. Cross-domain analysis further revealed reduced agreement between SHAP and attention-based explanations under domain shift, indicating lower explanation consistency across linguistically distinct domains. Qualitative error analysis further showed that implicit sentiment, sarcasm, and domain-specific slang remain major sources of prediction errors. Overall, the results demonstrate that superior predictive performance does not necessarily correspond to higher explanation faithfulness or stronger cross-domain stability. These findings highlight the importance of jointly evaluating predictive performance, explanation faithfulness, and explanation robustness when developing trustworthy transformer-based NLP systems.

Keywords: Attention-based explanations; Cross-domain evaluation; Explainable artificial intelligence (XAI); Explanation faithfulness; Hate speech detection; Natural language processing; SHAP; Transformer models.

Received: May, 23rd 2026
Revised: July, 4th 2026
Accepted: July, 5th 2026
Published: July, 21st 2026

1. Introduction

Transformer-based architectures such as BERT [1], RoBERTa [2], DistilBERT [3], and more recent variants such as DeBERTa [4] have become the dominant approaches in modern natural language processing (NLP) owing to their outstanding performance across a wide range of text classification tasks. Their widespread adoption in applications such as sentiment analysis, hate speech detection, and automated content moderation has further increased the need to understand how these models generate their predictions [5].

Despite their strong predictive performance, transformer models remain inherently difficult to interpret. This limitation becomes particularly critical in sensitive applications such as hate speech detection, where inaccurate or poorly justified predictions may have significant social and ethical consequences [6]. Regulatory frameworks, including the European Union AI Act, further emphasise the importance of transparency and accountability in high-risk AI systems. Consequently, explainability is increasingly regarded not only as an important research objective but also as a practical requirement for deploying trustworthy AI systems in real-world environments [7].

Existing explainability research in NLP generally follows two main paradigms. The first employs perturbation-based post-hoc methods, such as SHAP [8], to estimate the contribution of individual input tokens to model predictions. The second relies on transformer attention weights as indicators of token importance [9]. However, whether attention weights can reliably explain model behaviour remains an active topic of debate. Jain and Wallace [10] argued that attention weights often fail to provide faithful explanations, whereas Wiegreffe and Pinter [11] suggested that attention mechanisms can still offer meaningful interpretability under specific conditions. These contrasting findings highlight the need for more comprehensive empirical evaluations of explainability methods in transformer-based NLP systems.

Moreover, many existing explainability studies continue to rely primarily on qualitative interpretation, making it difficult to determine whether highlighted tokens genuinely influence model predictions or merely appear plausible to human observers. This distinction is particularly important in practical NLP applications where model predictions support sensitive decisions, such as content moderation and hate speech detection. Explanations that appear visually convincing may not accurately reflect the model's underlying decision-making process, potentially creating misplaced trust in automated systems. Therefore, quantitative faithfulness evaluation provides a more reliable basis for assessing whether explanation methods genuinely capture model behaviour rather than simply producing intuitive visualisations.

Another important challenge concerns the robustness of explainability methods under domain shift. Most previous studies evaluate interpretability using a single dataset or task setting [12], leaving it unclear whether explanation quality remains consistent across domains with substantially different linguistic characteristics. Long-form movie reviews and short-form social media posts differ considerably in lexical ambiguity, contextual dependency, and semantic structure [13]. These differences may influence both model predictions and token-level attribution patterns. As transformer-based NLP systems are increasingly deployed across heterogeneous environments, explanation methods that appear reliable within a single domain may become substantially less trustworthy when applied to linguistically distinct domains. This raises an important yet insufficiently explored question: Do explainability methods remain faithful when language distributions change across application domains?

Large-scale sentiment analysis datasets such as IMDb Movie Reviews [14], together with optimisation strategies such as AdamW regularisation [15], have become widely adopted benchmarks and training techniques in transformer-based NLP because of their effectiveness in improving model generalisation and training stability. Motivated by these research gaps, this study develops a unified experimental framework to jointly evaluate transformer architectures, explainability methods, and linguistic domains. Rather than proposing a new explainability algorithm, the study aims to systematically investigate the relationships among predictive performance, explanation faithfulness, and cross-domain consistency under substantially different language conditions. The main contributions of this study are summarised as follows:

- A comprehensive explainability evaluation of four representative transformer architectures (BERT, RoBERTa, DistilBERT, and DeBERTa) using two linguistically distinct datasets for sentiment analysis and hate speech detection.
- A quantitative faithfulness evaluation comparing SHAP and attention-based explanations as representative perturbation-based and intrinsic explainability paradigms through top-k token-removal analysis.
- A cross-domain consistency analysis using Spearman rank correlation to investigate explanation robustness under domain shift between formal long-form and informal short-form language.
- A qualitative investigation of linguistic phenomena associated with explanation failures, providing empirical evidence that predictive performance, explanation faithfulness, and explanation stability should be jointly considered when evaluating trustworthy NLP systems.

The remainder of this paper is organised as follows. Section 2 reviews the related work. Section 3 presents the proposed methodology. Section 4 reports the experimental results, qualitative analyses, and comparisons with related studies. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related Work

This section reviews previous studies related to transformer-based text classification, explainability methods in NLP, and quantitative evaluation of explanation faithfulness. The discussion is organised into three areas: transformer architectures for sentiment analysis and hate speech detection, explainability approaches based on SHAP and attention mechanisms, and recent advances in faithfulness evaluation and cross-domain explainability.

2.1. Transformer Models for Text Classification

Transformer-based architectures such as BERT [1], RoBERTa [2], DistilBERT [3], and DeBERTa [4] have become the dominant foundation models for a wide range of NLP tasks, including sentiment analysis, hate speech detection, question answering, and natural language inference. Their strong contextual representation capability and widespread adoption make them suitable candidates for investigating explanation behaviour across different linguistic domains.

Numerous studies have demonstrated the effectiveness of transformer models for text classification. In sentiment analysis, Tan et al. [16] combined RoBERTa embeddings with recurrent architectures, while Semary et al. [17] integrated transformer, convolutional, and recurrent networks to improve representation learning and classification performance. For hate speech detection, Mozafari et al. [6] showed that transformer-based models outperform conventional machine learning approaches despite challenges posed by sarcasm, implicit sentiment, and noisy social media language. Similarly, Sun et al. [18] demonstrated that reformulating aspect-based sentiment analysis as a sentence-pair classification task further improves transformer performance.

Recent studies have also evaluated newer transformer architectures beyond the original BERT family. Timoneda and Vallejo Vera [4] reported that DeBERTa consistently achieved competitive or superior performance across multiple political science text classification tasks owing to its disentangled attention mechanism. Likewise, Saputra et al. [19] demonstrated the robustness of DeBERTa for stance detection, suggesting that its effectiveness extends beyond sentiment analysis and hate speech detection.

Despite these advances, existing studies primarily evaluate transformer models from the perspective of predictive performance, while explanation quality receives considerably less attention. Furthermore, explainability is typically assessed within a single dataset or task, providing limited evidence regarding whether explanations remain faithful and consistent across linguistically distinct domains. These limitations motivate a systematic comparative evaluation of transformer architectures under cross-domain explainability settings.

2.2. Explainability in NLP

Explainability methods aim to improve the transparency of complex models by providing interpretable representations of their prediction process [7]. Among the most widely adopted post-hoc approaches is SHAP [8], which estimates the contribution of individual input features using Shapley values derived from cooperative game theory. By assigning additive feature attributions, SHAP quantifies the influence of each input token on the final prediction. Lundberg et al. [20] further demonstrated that SHAP often provides higher faithfulness than several gradient-based attribution methods across different model architectures.

Attention-based explanations represent an alternative paradigm by interpreting transformer attention weights as indicators of token importance [9]. However, their explanatory reliability remains controversial. Jain and Wallace [10] argued that different attention distributions can produce identical predictions, suggesting that attention weights do not necessarily provide faithful explanations. In contrast, Wiegreffe and Pinter [11] showed that attention mechanisms can still offer meaningful interpretability under specific evaluation settings, while Clark et al. [21] demonstrated that certain attention heads capture syntactic and semantic relationships. Consequently, this study compares SHAP and attention-based explanations because they represent two complementary explainability paradigms: SHAP provides perturbation-based feature attribution, whereas attention reflects the model's internal information flow.

Recent surveys on explainable NLP, including the work of Danilevsky et al. [12], identify explanation faithfulness and cross-domain generalisability as two important open challenges. Nevertheless, systematic quantitative comparisons between SHAP and attention-based

explanations under domain shift remain limited, particularly for transformer-based NLP systems. Therefore, this study focuses on these representative explainability paradigms to systematically evaluate explanation faithfulness and cross-domain consistency while minimising confounding factors introduced by comparing substantially different attribution algorithms.

2.3. Faithfulness Evaluation and Cross-Domain Explainability

Faithfulness evaluation aims to determine whether an explanation accurately reflects a model's underlying decision-making process rather than merely appearing plausible to human observers [22]. Jacovi and Goldberg [22] distinguished faithfulness from plausibility, arguing that explanations may seem convincing while failing to represent the actual reasoning process of the model. This distinction motivates the use of quantitative evaluation methods, such as token-removal analysis, to assess explanation quality objectively.

Ribeiro et al. [23] further demonstrated that NLP models achieving high predictive performance may still fail systematically on specific linguistic phenomena, including negation, temporal reasoning, and entity handling. These findings suggest that predictive performance alone is insufficient for evaluating model reliability. Consequently, explanation methods should be assessed not only by visual interpretability but also by their consistency with the model's underlying decision process.

In the context of hate speech detection, Böck et al. [24] compared gradient-based, perturbation-based, attention-based, and prototype-based explanation methods, reporting that perturbation-based approaches such as SHAP generally produced more reliable explanations than attention-based alternatives. Nevertheless, most existing studies evaluate explanation quality within a single dataset or task, leaving the robustness of explainability methods under domain shift largely unexplored.

Recent studies have further emphasised the need for more rigorous explainability evaluation in transformer-based NLP systems. Wiegreffe and Pinter [11] argued that attention mechanisms may still provide useful explanatory signals under specific evaluation settings, whereas Atanasova et al. [25] showed that explanation methods often fail to capture deeper linguistic reasoning despite producing visually plausible token attributions. Similarly, Hase and Bansal [26] highlighted the importance of quantitative evaluation frameworks capable of measuring explanation faithfulness and consistency under varying input conditions.

Table 1 summarises representative studies and highlights the remaining limitations addressed in this work. Although previous research has investigated transformer models or explainability methods individually, no prior study has jointly compared multiple transformer architectures using both SHAP and attention-based explanations under cross-domain settings with quantitative faithfulness and consistency evaluation.

Table 1. Comparison of explainability evaluation strategies in related studies

Study	Models	SHAP	Attention	Quantitative Evaluation	Cross-domain Analysis
Sun et al. [18]	BERT	-	-	-	-
Mozafari et al. [6]	BERT	-	-	-	-
Tan et al. [16]	RoBERTa	-	-	-	-
Jain & Wallace [10]	RNN	-	✓	-	-
Böck et al. [24]	BERT	✓	✓	-	-
This study	BERT, RoBERTa, DistilBERT, DeBERTa	✓	✓	✓	✓

2.4. Research Gap

Table 1 highlights several important limitations in existing explainability studies for transformer-based NLP models. Previous work has primarily focused on improving predictive performance or providing qualitative interpretations within individual datasets. Although SHAP and attention-based explanations have each been studied extensively, systematic quantitative comparisons between these explainability paradigms remain limited. Furthermore, most studies evaluate explanation quality only under in-domain settings, leaving it unclear

whether explanation behaviour remains faithful and consistent across linguistically distinct domains.

Another important limitation concerns the relationship between predictive performance and explanation reliability. Existing transformer studies generally report classification accuracy as the primary evaluation criterion, whereas quantitative explainability evaluation receives considerably less attention. Consequently, it remains uncertain whether models achieving superior predictive performance also produce explanations that are more faithful and robust under domain shift.

To address these gaps, this study develops a unified evaluation framework that systematically compares four transformer architectures using two complementary explainability methods across formal and informal language domains. Rather than proposing a new explainability algorithm, the study aims to provide empirical evidence on the relationships among predictive performance, explanation faithfulness, and cross-domain consistency in transformer-based NLP systems.

3. Proposed Method

This section describes the proposed cross-domain explainability evaluation framework. The framework integrates transformer-based text classification, complementary explainability methods, quantitative faithfulness evaluation, cross-domain consistency analysis, and qualitative error analysis into a unified experimental pipeline. Rather than introducing a new explainability algorithm, the proposed framework provides a systematic methodology for comparing explanation quality across multiple transformer architectures and linguistically distinct domains. The overall workflow of the proposed framework is illustrated in Figure 1.

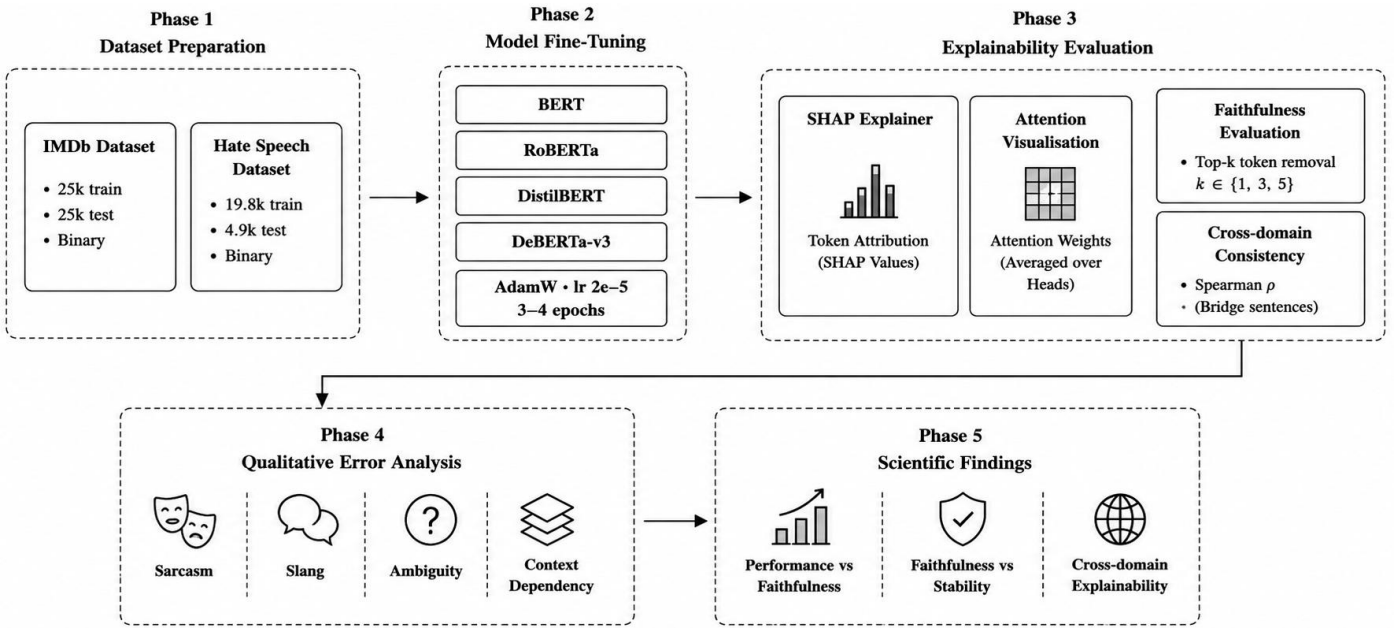

Figure 1. Proposed cross-domain explainability evaluation framework for transformer-based NLP classification using SHAP and attention-based attribution methods.

3.1. Datasets

Two publicly available benchmark datasets were selected to represent substantially different linguistic environments for evaluating explanation faithfulness and cross-domain consistency. Rather than simply comparing model performance across two NLP tasks, the selected datasets were intentionally chosen to create a realistic domain-shift scenario in which transformer models operate under contrasting language characteristics. This design enables investigation of whether explanation methods remain faithful and stable when applied to linguistically distinct domains.

The first dataset is IMDb Movie Reviews [14], which contains 50,000 movie reviews evenly divided into 25,000 training and 25,000 test samples with binary sentiment labels (positive/negative). Reviews are generally long (mean: 231 tokens), grammatically well structured,

and linguistically diverse, making the dataset suitable for evaluating explainability in relatively clean and formal language settings.

The second dataset is Hate Speech Offensive [13], which consists of 24,783 Twitter posts annotated into three classes: hate speech, offensive language, and neither. Following prior work [6], the labels are binarised by treating hate speech as the positive class while merging offensive and neutral content into the negative class. Because no official train–test split is provided, a stratified 80/20 split is adopted. Compared with IMDb, this dataset presents substantially greater challenges due to shorter texts (mean: 21 tokens), informal writing styles, frequent slang and abbreviations, higher lexical ambiguity, and severe class imbalance, with hate speech representing only 5.8% of the total samples.

The contrasting characteristics of both datasets are summarised in Table 2. Compared with IMDb, the Hate Speech dataset contains noisier and more context-dependent language with lower discourse coherence and substantially higher lexical ambiguity. These characteristics increase semantic uncertainty and make token-level attribution more challenging, thereby providing a more demanding benchmark for evaluating explanation faithfulness and cross-domain consistency.

Table 2. Comparison of the datasets used in this study.

Property	IMDb	Hate Speech
Total samples	50,000	24,783
Training samples	25,000	19,826
Test samples	25,000	4,957
Label type	Binary (pos/neg)	Binary (hate/non-hate)
Mean token length	231	21
Minority class (%)	50.0	5.8
Source domain	Movie reviews	Twitter/X social media
Context availability	High (complete reviews)	Low (short isolated posts)
Lexical ambiguity	Low–Moderate	High
Discourse coherence	High	Low
Noise level	Low	High
Slang usage	Rare	Frequent
Text type	Formal long-form	Informal short-form
Linguistic characteristics	Formal, structured	Slang, abbreviations

The two datasets were deliberately selected because they represent substantially different linguistic conditions. IMDb reviews consist of long-form, grammatically structured text with explicit sentiment expressions and coherent discourse patterns, whereas the Hate Speech dataset contains short-form social media content characterised by informal language, slang, abbreviations, lexical ambiguity, and highly context-dependent expressions. These differences create a realistic domain-shift setting in which explanation methods must operate across contrasting language conditions.

The substantial differences in text length, contextual dependency, linguistic structure, and class distribution provide an appropriate testbed for assessing whether explanation quality remains faithful and consistent under varying levels of language complexity and noise. This evaluation setting reflects practical NLP deployment scenarios, where models are frequently applied beyond the distribution of the data used during training. Accordingly, this study hypothesises that explanation methods exhibiting strong faithfulness in the relatively clean and structured IMDb domain may demonstrate reduced faithfulness and cross-domain consistency when applied to the more linguistically challenging Hate Speech domain.

3.2. Pre-trained Models

Four pre-trained transformer architectures with different model capacities and attention formulations were selected to investigate whether architectural differences influence explanation faithfulness and cross-domain consistency. Rather than identifying the best-performing classifier, the selected models provide representative transformer variants with varying

representation capabilities while remaining comparable in terms of fine-tuning strategy and evaluation protocol.

BERT [1], RoBERTa [2], dan DistilBERT [3] represent widely adopted transformer architectures based on conventional self-attention mechanisms. In contrast, DeBERTa [4] employs a disentangled attention mechanism that separately models content and positional information during contextual representation learning. This architectural variation enables a systematic comparison of whether different attention formulations influence explanation behaviour under linguistically distinct domains. Furthermore, all four models have demonstrated strong performance across a wide range of NLP applications, including sentiment analysis, hate speech detection, stance detection, question answering, natural language inference, and political text classification [1]–[4], [19], making them suitable candidates for comparative explainability evaluation. The characteristics of the evaluated transformer models are summarised in Table 3.

Table 3. Pre-trained transformer models evaluated in this study.

Model	HuggingFace ID	Params	Layers	Pre-training Data
BERT [1]	bert-base-uncased	110M	12	Wikipedia + BookCorpus
RoBERTa [2]	roberta-base	125M	12	Wikipedia + CC-News (160 GB)
DistilBERT [3]	distilbert-base-uncased	66M	6	Distilled from BERT
DeBERTa [4]	microsoft/deberta-v3-base	184M	12	Large-scale web and book corpora

To ensure a fair comparison, all transformer models were fine-tuned using an identical training protocol. The AdamW optimiser [15] was adopted with a learning rate of 2×10^{-5}, a training and evaluation batch size of 8, and a linear learning-rate scheduler with a warmup ratio of 0.1. The maximum sequence length was set to 512 tokens for IMDb and 128 tokens for Hate Speech to reflect the substantially different average text lengths of the two datasets. The best-performing checkpoint was selected according to the highest validation F1-score.

Training was conducted for three epochs on IMDb and four epochs on Hate Speech. The larger IMDb dataset converged more rapidly because of its greater number of training samples and relatively cleaner linguistic structure, whereas the Hate Speech dataset required an additional epoch owing to its shorter, noisier, and more heterogeneous language characteristics. Applying the same optimisation strategy across all models ensures that observed differences in classification performance and explanation quality primarily reflect differences in transformer architectures rather than variations in training configuration. The complete fine-tuning configuration is presented in Table 4.

Table 4. Hyperparameter configuration for fine-tuning experiments.

Property	IMDb	Hate Speech
Learning rate	2×10^{-5}	2×10^{-5}
Batch size (train/eval)	8 / 8	8 / 8
Epochs	3	4
Maximum sequence length	512 tokens	128 tokens
Optimiser	AdamW	AdamW
Learning-rate scheduler	Linear with warmup	Linear with warmup
Warmup ratio	0.1	0.1
Random seed	42	42

3.3. Explainability Methods

This study employs two complementary explainability methods to analyse transformer predictions: SHAP and attention-based explanations. Rather than comparing numerous attribution techniques, the objective is to evaluate two representative explainability paradigms within a unified experimental framework. SHAP represents a perturbation-based post-hoc approach, whereas attention-based explanations derive token importance directly from the transformer's internal attention distributions.

The two methods were selected because they offer complementary perspectives on model interpretability while remaining among the most widely adopted explainability approaches in transformer-based NLP. SHAP estimates the contribution of individual tokens to the final prediction, whereas attention-based explanations reflect the model's internal information flow. Applying both methods to the same models, datasets, and evaluation protocol enables differences in explanation quality to be attributed primarily to the explainability strategy rather than to variations in model architecture or experimental settings. Accordingly, the comparison focuses on the contrast between perturbation-based and intrinsic explanation strategies for evaluating explanation faithfulness and cross-domain consistency.

3.3.1. SHapley Additive exPlanations (SHAP)

SHAP [8] was selected as the representative perturbation-based explainability method because it provides theoretically grounded token-level feature attribution based on cooperative game theory. For a model with n input features, the Shapley value of token i is computed as:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!}[f(S \cup \{i\}) - f(S)] \quad (1)$$

where N denotes the complete feature set and S represents a subset excluding feature i.

Within the proposed framework, SHAP is used to generate token-level attribution scores for faithfulness evaluation, cross-domain consistency analysis, and qualitative explanation comparison. Explanations are generated using the unified shap.Explainer interface with the corresponding HuggingFace tokenizer as the text masker. Each input is tokenised using automatic padding and truncation, while prediction probabilities are obtained from the model logits through the softmax function.

To ensure a fair comparison across all transformer architectures, the same tokenizer, masking strategy, prediction function, and preprocessing settings are applied throughout the experiments without model-specific parameter tuning. Global explanation scores are computed from 100 randomly selected test instances to reduce computational cost, whereas local explanations are generated for representative samples used in the qualitative analysis. This consistent configuration ensures that differences in attribution quality primarily reflect differences in transformer representations rather than variations in the explainability setup.

3.3.2. Attention Visualisation

Attention-based explanations were selected as the representative intrinsic explainability method because they estimate token importance directly from the transformer's internal attention distributions without requiring additional perturbation or surrogate models. Unlike SHAP, which measures the contribution of each token to the prediction outcome, attention visualisation reflects how contextual information is distributed within the model. Evaluating both approaches under the same experimental framework enables a systematic comparison between perturbation-based and intrinsic explanation strategies. Attention scores are extracted from the final transformer layer and averaged across all H attention heads [9], [21]:

$$A_{\text{token}} = \frac{1}{H}\sum_{h=1}^{H} A^{(h)}_{[\text{CLS}],\,\text{token}} \quad (2)$$

where $A^{(h)}_{[\text{CLS}],\,\text{token}}$ denotes the attention weight from the $[CLS]$ token to each input token in attention head h.

The final transformer layer is selected because it captures task-specific contextual representations most closely associated with the final classification decision [21]. Attention weights are averaged across all heads to obtain a single and consistent token-importance distribution while reducing head-specific variability across transformer architectures. The same extraction strategy is applied to all evaluated models to ensure methodological consistency throughout the experiments. Consequently, differences in explanation quality primarily reflect model behaviour rather than variations in attention extraction.

Although attention weights were not originally designed as explanation mechanisms and their faithfulness remains actively debated [10], [11], they provide an informative baseline for

comparison with perturbation-based explanations. This complementary evaluation enables a more comprehensive assessment of explanation faithfulness and cross-domain consistency.

3.4. Faithfulness Evaluation

The primary objective of the proposed evaluation framework is to determine whether an explanation correctly identifies the tokens that genuinely influence model predictions. To quantify explanation faithfulness, a top-k token-removal strategy is adopted following Jacovi and Goldberg [22]. For each test instance, the k highest-ranked tokens identified by an explanation method are replaced with the tokenizer's [MASK] token before the modified input is re-evaluated by the fine-tuned model.

Mask-based token replacement is adopted because it suppresses the contribution of the selected tokens while preserving the original sentence length and positional structure. The resulting faithfulness score is defined as the mean absolute change in prediction confidence:

$$\text{Faithfulness}_k = \frac{1}{N}\sum_{i=1}^{N}\left|f(x_i) - f\left(x_i^{(k)}\right)\right| \tag{3}$$

where $f(x_i)$ represents the original prediction confidence and $f(x_i^{(k)})$ represents the prediction confidence after removing the top-k attributed tokens.

Higher faithfulness scores indicate that the identified tokens exert greater influence on the model prediction. Accordingly, larger confidence degradation corresponds to higher explanation faithfulness. Faithfulness is evaluated using $k \in \{1,3,5\}$ on 200 randomly selected test instances from each dataset. These values progressively assess the influence of the most important token ($k = 1$) and groups of highly ranked tokens ($k = 3$ and $k = 5$). Restricting token removal to a small number of tokens preserves the overall semantic structure of the input while introducing meaningful perturbations for quantitative evaluation.

3.5. Cross-Domain Consistency Analysis

Cross-domain consistency analysis is performed to evaluate whether explanation methods produce stable token rankings when transformer models are applied to linguistically distinct domains. To ensure a fair comparison, all fine-tuned models are evaluated using the same set of 50 manually constructed bridge sentences. These short opinion-like sentences are intentionally designed to be applicable across both sentiment analysis and hate speech domains while avoiding explicit sentiment polarity, hate-related expressions, and domain-specific lexical cues that could bias explanation rankings.

The bridge sentences contain between 6 and 14 tokens and were independently reviewed by the authors to ensure semantic neutrality and linguistic consistency. A fixed sentence set is adopted for all transformer models to ensure that differences in explanation consistency arise from model behaviour rather than variations in input text. A representative example is "This situation seems difficult to understand." The complete bridge sentence set is provided as supplementary material to support transparency and reproducibility. Cross-domain consistency is quantified using Spearman's rank correlation (ρ) between the token rankings generated by SHAP and attention-based explanations:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4}$$

where d_i denotes the rank difference for token i, and n is the total number of tokens. Higher correlation values indicate stronger agreement between explanation methods, whereas lower values reflect reduced explanation consistency under domain shift. The metric is computed separately for models fine-tuned on IMDb and Hate Speech to isolate the effect of linguistic domain on explanation stability.

3.6. Experimental Procedure

The complete evaluation procedure is summarised in Algorithm 1. Each transformer model is independently fine-tuned on the IMDb and Hate Speech datasets using an identical training configuration. After classification performance is evaluated, SHAP and attention-based explanations are generated using the corresponding explainability configuration

described in the previous subsections. Faithfulness is then quantified through top-k token-removal analysis, followed by cross-domain consistency analysis using the shared bridge sentence set. Finally, qualitative error analysis is conducted to investigate explanation behaviour on representative misclassified samples. The resulting classification, faithfulness, and consistency metrics are subsequently compared across all transformer architectures and datasets.

Algorithm 1. Cross-Domain Explainability Evaluation Procedure

INPUT: Datase $D_{\text{IMDb}}, D_{\text{Hate}}$;Models $\mathcal{M} = \{\text{BERT, RoBERTa, DistilBERT, DeBERTa}\}$
OUTPUT: Classification metrics, faithfulness scores, and cross-domain consistency results

1: for all $m \in \mathcal{M}$ do
2: Fine-tune $m$on D_{IMDb}; evaluate classification performance metrics (Accuracy, Macro F1, Precision, and Recall)
3: Fine-tune $m$on D_{Hate}; evaluate classification performance metrics (Accuracy, Macro F1, Precision, and Recall)
4: Compute SHAP values (Eq. 1) on 200 test samples per dataset
5: Extract attention weights (Eq. 2), averaged across all heads in the final transformer layer
6: Compute Faithfulness_k(Eq. 3) at $k \in \{1,3,5\}$
7: Compute Spearman's ρ(Eq. 4) on 50 bridge sentences per model variant
8: Perform qualitative error analysis on 150 randomly sampled misclassified instances per dataset through manual review by the authors
9: end for
10: Aggregate and compare results across all models and domains

4. Results and Discussion

All experiments were conducted on a single NVIDIA A100 GPU (40 GB VRAM). Fine-tuning was implemented using PyTorch 2.0 with the HuggingFace Transformers 4.38 framework [5], while SHAP library 0.44 [8] was used for attribution analysis. Reproducibility was ensured by fixing the random seed to 42 across NumPy, PyTorch, and Python. Training converged consistently within the allocated epochs without signs of instability or early divergence. Each model required approximately 45 minutes to train on IMDb and 12 minutes on Hate Speech. The best checkpoint was selected according to the highest validation F1-score.

4.1. Classification Performance

Table 5 summarises the classification performance of the four transformer architectures on both datasets. Overall, DeBERTa achieves the best results across all evaluation metrics. On IMDb, it reaches 95.6% accuracy and 95.4% F1-score, outperforming RoBERTa by 1.4 percentage points and BERT by 2.8 percentage points in accuracy. On the Hate Speech dataset, DeBERTa again achieves the highest performance with 91.3% accuracy and 89.8% F1-score, exceeding RoBERTa by 1.6 percentage points and BERT by 3.9 percentage points. These results suggest that DeBERTa's disentangled attention mechanism and enhanced positional encoding improve contextual representation learning, particularly for linguistically challenging text.

Although all evaluated models achieve competitive classification performance, accuracy alone does not fully reflect explanation quality. The following analyses therefore investigate whether superior predictive performance is accompanied by stronger explanation faithfulness and greater cross-domain stability.

Performance consistently decreases when models are transferred from IMDb to the Hate Speech dataset. As shown in Table 5, DeBERTa experiences the smallest reduction in accuracy (4.3 percentage points), followed by RoBERTa (4.5), BERT (5.4), and DistilBERT (6.2). The smaller degradation observed for DeBERTa indicates greater robustness under domain shift, whereas the larger decline of DistilBERT suggests that parameter reduction may limit its ability to capture subtle semantic dependencies in noisy and context-dependent social media language. This observation is consistent with previous studies reporting that hate speech detection remains considerably more challenging because of implicit expressions, linguistic ambiguity, and substantial lexical variation across online communities [27], [28].

Table 5. Classification performance across datasets

Dataset	Model	Accuracy	F1-Score	Precision	Recall
IMDb	BERT	92.8	92.6	93.1	92.2
	RoBERTa	94.2	94.0	94.5	93.6
	DistilBERT	91.3	91.1	91.8	90.5
	DeBERTa	95.6	95.4	95.9	95.0
Hate Speech	BERT	87.4	85.9	86.3	85.5
	RoBERTa	89.7	88.1	88.6	87.7
	DistilBERT	85.1	83.4	84.0	82.9
	DeBERTa	91.3	89.8	90.2	89.4

The classification results establish the predictive performance baseline used in the subsequent explainability evaluation. The following subsection examines whether the most accurate models also produce the most faithful explanations.

Table 6. Average confidence degradation (%) after top-k token removal across all evaluated transformer architectures. Higher values indicate stronger explanation faithfulness.

Dataset	Method	k=1	k=3	k=5
IMDb	BERT–SHAP	18.4	31.2	42.7
	BERT–Attention	11.3	19.8	28.4
	RoBERTa–SHAP	21.3	35.8	47.1
	RoBERTa–Attention	13.2	22.4	31.7
	DistilBERT–SHAP	16.9	28.4	38.5
	DistilBERT–Attention	10.8	18.3	26.1
	DeBERTa–SHAP	22.8	37.6	49.3
	DeBERTa–Attention	14.1	23.8	33.2
Hate Speech	BERT–SHAP	14.2	24.1	33.8
	BERT–Attention	8.7	14.9	21.3
	RoBERTa–SHAP	16.8	27.3	37.4
	RoBERTa–Attention	9.4	16.2	23.5
	DistilBERT–SHAP	12.6	21.5	30.2
	DistilBERT–Attention	7.9	13.4	19.7
	DeBERTa–SHAP	18.1	29.0	39.8
	DeBERTa–Attention	10.2	17.5	25.1

4.2. Faithfulness Evaluation

Beyond predictive performance, explanation methods should identify the tokens that genuinely influence model predictions. To evaluate this capability, top-k token-removal analysis is performed to compare the faithfulness of SHAP and attention-based explanations across both datasets. Larger confidence degradation indicates higher explanation faithfulness because removing the identified tokens causes a greater reduction in prediction confidence. The complete results are summarised in Table 6, while Figure 2 illustrates the overall trend across different values of k.

Across all transformer architectures, SHAP consistently produces substantially larger confidence degradation than attention-based explanations for both datasets, indicating stronger explanation faithfulness regardless of model architecture. A consistent performance hierarchy is also observed, with DeBERTa achieving the highest faithfulness scores, followed by RoBERTa, BERT, and DistilBERT. This pattern suggests that richer contextual representations generally improve the identification of prediction-relevant tokens, although the relatively small differences between adjacent architectures indicate that explanation faithfulness depends not only on model capacity but also on the quality of contextual representations learned during fine-tuning.

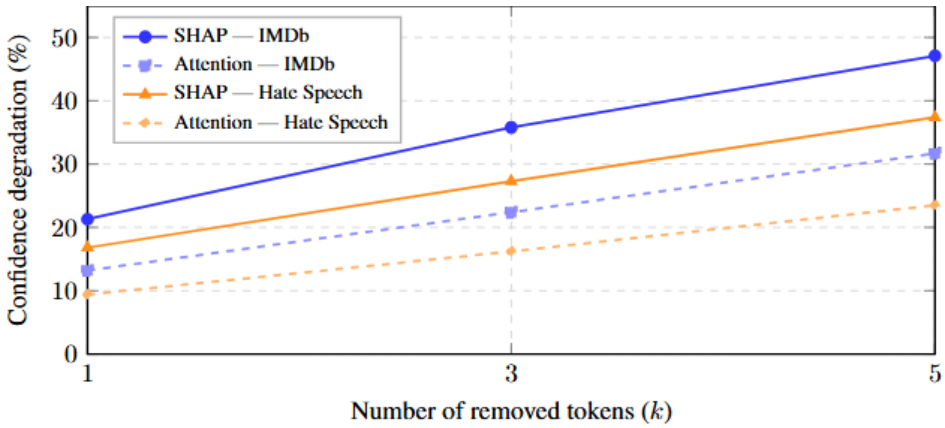

Figure 2. Faithfulness comparison between SHAP and attention-based explanations across increasing values of k.

The performance gap between SHAP and attention-based explanations becomes more pronounced as k increases, indicating that SHAP captures broader sets of semantically informative tokens. This difference reflects the underlying attribution mechanisms of the two approaches. SHAP directly measures changes in prediction confidence after token perturbation, producing explanations that are explicitly linked to model behaviour. In contrast, attention weights primarily describe the distribution of contextual information and are not explicitly optimised to represent the causal contribution of individual tokens to the final prediction. Consequently, attention may highlight linguistically prominent tokens without necessarily identifying those that most strongly influence the model output.

Across both explanation methods, confidence degradation is consistently lower on the Hate Speech dataset than on IMDb. This finding is closely related to the linguistic characteristics summarised in Table 2. Unlike the grammatically structured IMDb reviews containing explicit sentiment expressions, hate speech posts frequently rely on slang, abbreviations, implicit meaning, and context-dependent language. As a result, prediction-relevant information tends to be distributed across multiple contextual cues rather than concentrated in a few highly influential tokens. Removing only a small number of highly ranked tokens therefore produces smaller confidence changes, indicating that explanation faithfulness is influenced not only by the explanation method itself but also by the linguistic characteristics of the target domain.

From a practical perspective, these findings emphasise the importance of evaluating explanation faithfulness rather than relying solely on visual interpretability. Explanations that appear intuitively plausible may not accurately reflect the features responsible for model predictions. This distinction is particularly important in high-risk applications such as hate speech detection, where misleading explanations may reduce transparency, accountability, and trust in automated decision-making systems.

4.3. Cross-Domain Consistency Analysis

While the faithfulness analysis demonstrates that SHAP consistently identifies prediction-relevant tokens more effectively than attention-based explanations, cross-domain consistency analysis evaluates whether these explanations remain stable when models are applied across linguistically distinct domains. Spearman's rank correlation is computed between the token rankings generated by SHAP and attention-based explanations using the shared bridge sentence set. The average consistency scores are summarised in Table 7 and visualised in Figure 3.

Across all evaluated transformer architectures, SHAP consistently achieves higher cross-domain consistency than attention-based explanations. DistilBERT records the highest SHAP consistency ($\rho = 0.64$), followed by BERT ($\rho = 0.61$) and RoBERTa ($\rho = 0.58$), whereas attention-based explanations produce substantially lower agreement across all models. These results indicate that perturbation-based attribution remains more stable than attention-derived explanations when linguistic distributions change across domains.

The consistency gap ($\Delta\rho$) reported in Table 7 further highlights the divergence between the two explanation methods. RoBERTa and DeBERTa exhibit the largest gaps ($\Delta\rho = 0.19$), while BERT and DistilBERT show slightly smaller differences ($\Delta\rho = 0.17$). This trend suggests that attention-based explanations become increasingly sensitive to domain-specific contextual representations as transformer architectures learn richer contextual features.

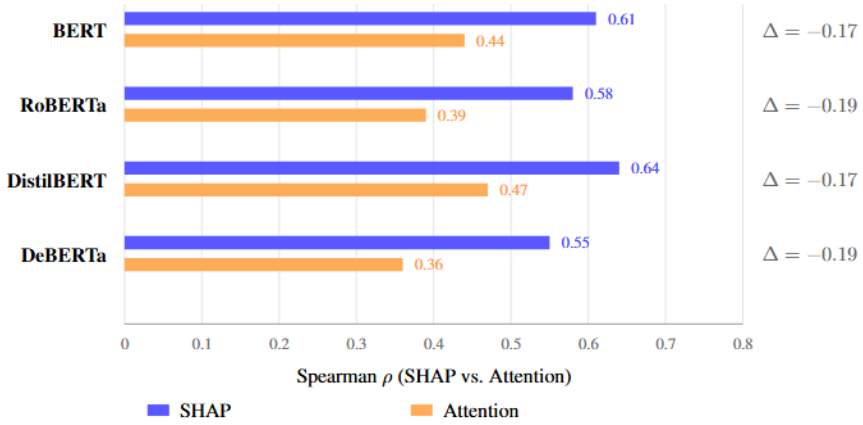

Figure 3. Cross-domain consistency comparison between SHAP and attention-based explanations using Spearman's rank correlation.

Table 7. Cross-domain consistency comparison using Spearman's rank correlation (ρ) and consistency gap ($\Delta\rho$) between SHAP and attention-based explanations.

Model	SHAP	Attention	Gap ($\Delta\rho$)
BERT	0.61	0.44	0.17
RoBERTa	0.58	0.39	0.19
DistilBERT	0.64	0.47	0.17
DeBERTa	0.55	0.36	0.19

Interestingly, DistilBERT achieves the highest explanation consistency despite having the lowest classification performance among the evaluated models. Conversely, DeBERTa, although consistently delivering the highest predictive performance, records the lowest cross-domain consistency. This finding suggests that richer contextual representations improve classification capability but do not necessarily produce more stable explanations under domain shift. One possible explanation is that larger transformer architectures exploit increasingly domain-specific contextual cues, making token-importance rankings more sensitive to changes in linguistic distributions.

Dataset characteristics further support this observation. Compared with the grammatically structured IMDb reviews, the Hate Speech dataset contains shorter texts, greater lexical ambiguity, frequent slang, and stronger contextual dependency. These characteristics increase semantic uncertainty and make explanation methods more sensitive to domain-specific language variation, particularly for attention-based explanations. In contrast, SHAP maintains relatively stable token rankings because its perturbation-based attribution directly measures the influence of individual tokens on prediction confidence. Overall, the results demonstrate that explanation stability represents an independent dimension of trustworthy NLP evaluation rather than a direct consequence of predictive performance. Therefore, predictive performance, explanation faithfulness, and cross-domain consistency should be considered complementary evaluation criteria when selecting transformer models for deployment in heterogeneous linguistic environments.

4.4. Qualitative Explanation Examples

While the quantitative faithfulness analysis provides objective evidence of explanation quality, it does not fully illustrate how different explanation methods behave at the token level. Therefore, qualitative examples are presented to compare the attribution patterns generated by SHAP and attention-based explanations on representative samples from both

datasets. Figure 4 shows token-level explanations for sentiment analysis and hate speech detection. The example sentences are lightly simplified for readability while preserving the semantic characteristics relevant to the original model predictions.

In the IMDb example, SHAP assigns the highest importance to sentiment-bearing expressions such as "poorly written" and "unsatisfying", which directly convey negative sentiment. In contrast, attention-based explanations distribute importance across more general contextual tokens, including "movie" and "was", whose contributions to sentiment polarity are less direct.

A similar pattern is observed for the Hate Speech sample. SHAP consistently highlights explicit exclusionary expressions such as "go back" and "nobody wants", whereas attention-based explanations assign importance more broadly across surrounding contextual tokens. These examples indicate that perturbation-based attribution more effectively identifies tokens that directly influence model predictions.

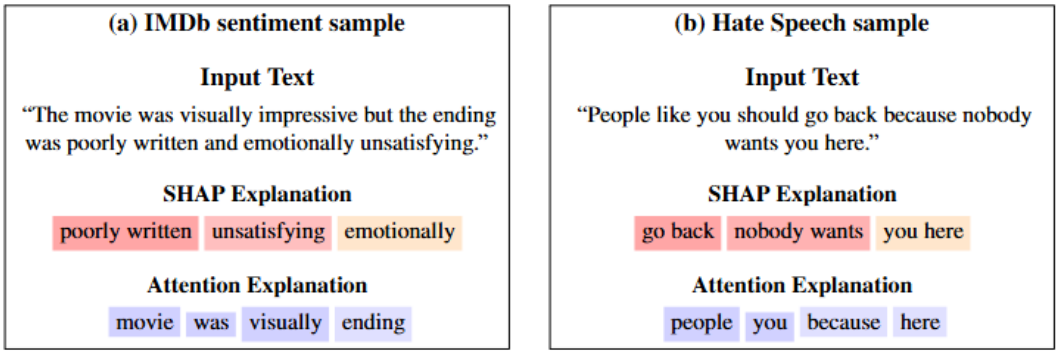

Figure 4. Representative comparison of SHAP and attention-based token attributions for sentiment analysis and hate speech detection.

Overall, the qualitative observations are consistent with the quantitative results presented in Section 4.2. Across both datasets, SHAP focuses on semantically meaningful tokens that contribute directly to prediction confidence, whereas attention-based explanations produce more diffuse attribution patterns. This behaviour explains the larger confidence degradation observed for SHAP during the faithfulness evaluation. These examples further demonstrate that visually plausible explanations are not necessarily faithful explanations. Although attention-based explanations often emphasise linguistically prominent tokens, such tokens do not always correspond to those that have the greatest influence on the model's decision. Consequently, the observed differences between SHAP and attention-based explanations reflect genuine differences in explanation faithfulness rather than merely alternative visualisation styles.

4.5. Qualitative Error Analysis

To better understand the limitations of transformer explanations, qualitative error analysis was conducted on 150 randomly selected misclassified instances from each dataset. The errors were manually grouped into recurring linguistic categories, including implicit sentiment, sarcasm or irony, ambiguous phrasing, domain-specific slang, and limited contextual information. Representative examples are presented in Table 8, while the distribution of error categories is summarised in Figure 5.

For the IMDb dataset, implicit sentiment and sarcasm account for the largest proportion of misclassified samples. Transformer models frequently struggle with reviews containing mixed sentiment, indirect emotional expressions, or context-dependent evaluations that require broader discourse understanding beyond individual sentiment-bearing words.

In the Hate Speech dataset, domain-specific slang and ambiguous expressions represent the dominant error sources. Many posts rely on implicit or context-dependent offensive language that cannot be interpreted reliably without additional conversational, cultural, or pragmatic context. The short and informal nature of social media text further increases lexical ambiguity, making both classification and explanation more challenging.

Table 8. Representative examples of misclassified samples and dominant error characteristics.

Dataset	Sample Text	Error Category	Prediction
IMDb	"The acting was great, but somehow the movie still felt emotionally empty."	Implicit sentiment	Positive
IMDb	"Amazing movie if your goal is to fall asleep after ten minutes."	Sarcasm / irony	Positive
Hate Speech	"People like you should just disappear from this country."	Implicit hate expression	Neutral
Hate Speech	"Bruh these clowns are ruining everything lol."	Domain-specific slang	Neutral

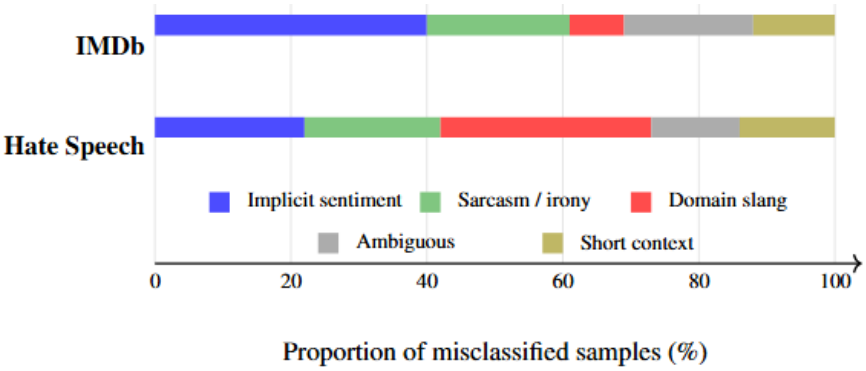

Figure 5. Distribution of linguistic error categories across the IMDb and Hate Speech datasets.

Interestingly, many misclassified samples exhibit less concentrated SHAP attribution patterns together with more diffuse attention distributions, suggesting that explanation uncertainty increases when models encounter linguistically ambiguous inputs. This behaviour is particularly evident for expressions involving implicit meaning, sarcasm, and domain-specific slang, where prediction-relevant information is distributed across multiple contextual cues rather than a small set of salient tokens. Taken together, the qualitative error analysis complements the quantitative findings presented in Sections 4.2 and 4.3. The results indicate that explanation faithfulness and cross-domain consistency are influenced not only by the explainability method but also by the linguistic characteristics of the target domain. Models operating on noisy, context-dependent language produce explanations that are generally less stable and more sensitive to linguistic variation, reinforcing the importance of evaluating explanation robustness alongside predictive performance.

4.6. Comparison with Related Studies

Previous studies on transformer-based NLP have primarily focused on improving predictive performance for specific tasks, such as sentiment analysis and hate speech detection, while explainability has often been evaluated qualitatively or within a single dataset [6]10,16,18,23]. Consequently, existing studies provide limited evidence regarding whether explanations remain faithful and consistent when transformer models are applied across linguistically different domains.

Rather than serving as a direct performance benchmark, the comparison presented in Table 9 highlights differences in research scope and explainability evaluation. Since the datasets, experimental settings, and model configurations differ across studies, the reported classification accuracies should be interpreted within their respective experimental contexts rather than compared directly.

Compared with previous studies, the primary contribution of the proposed framework lies in its evaluation strategy rather than in achieving the highest classification performance. The framework jointly evaluates predictive performance, explanation faithfulness, and cross-domain consistency using multiple transformer architectures under a unified experimental protocol. This enables a more comprehensive assessment of explanation robustness than studies that evaluate attribution quality within a single dataset or rely solely on qualitative interpretation.

The results further indicate that explanation quality cannot be inferred solely from predictive performance. While DeBERTa achieves the highest classification accuracy, SHAP consistently provides more faithful explanations, and cross-domain analysis reveals that explanation stability varies across transformer architectures. These observations demonstrate that predictive performance, explanation faithfulness, and cross-domain consistency represent complementary aspects of model evaluation.

Table 9. Comparison with related studies on transformer-based NLP classification and explainability evaluation.

Reference	Method	Dataset	Acc.	Explainability Evaluation
Sun et al. [18]	BERT + auxiliary sentence	ABSA	95.1	None
Mozafari et al. [6]	BERT + transfer learning	Hate Speech	88.9	None
Tan et al. [16]	RoBERTa–LSTM hybrid	IMDb	92.4	None
Jain and Wallace [10]	RNN + attention analysis	Multiple	–	Qualitative
Böck et al. [24]	Transformer + multi-XAI	Hate Speech	–	Qualitative
This study	BERT, RoBERTa, DistilBERT, DeBERTa + SHAP + Attention	IMDb + Hate Speech	95.6	Quantitative faithfulness and cross-domain evaluation

Taken together, the comparison with previous studies suggests that trustworthy evaluation of transformer-based NLP systems should extend beyond conventional classification metrics. Incorporating quantitative faithfulness analysis together with cross-domain consistency provides a broader perspective on explanation reliability and offers a practical framework for evaluating transformer models intended for deployment across heterogeneous linguistic environments.

5. Conclusions

This study evaluated the explainability of four transformer architectures (BERT, RoBERTa, DistilBERT, and DeBERTa) across two linguistically distinct NLP domains using a unified evaluation framework based on explanation faithfulness and cross-domain consistency. Rather than proposing a new explainability method, this study investigated how predictive performance, explanation faithfulness, and explanation stability interact under domain shift.

Four main findings emerged from this study. First, DeBERTa consistently achieved the highest classification performance but did not exhibit the strongest cross-domain explanation consistency, indicating that predictive performance does not necessarily translate into more robust explanations. Second, SHAP consistently produced higher faithfulness scores than attention-based explanations across all evaluated models and datasets, demonstrating that perturbation-based explanations better capture prediction-relevant features. Third, cross-domain evaluation showed that explanation consistency varies across transformer architectures and decreases under domain shift, highlighting explanation stability as an independent evaluation dimension. Fourth, qualitative error analysis revealed that linguistic characteristics, including contextual dependency, lexical ambiguity, sarcasm, and informal language, substantially influence explanation faithfulness and stability.

Taken together, these findings demonstrate that predictive performance, explanation faithfulness, and cross-domain consistency should be regarded as complementary criteria for evaluating trustworthy transformer-based NLP systems. By integrating quantitative faithfulness analysis with cross-domain consistency evaluation, the proposed framework provides a more comprehensive perspective on explanation robustness across heterogeneous linguistic domains. At the same time, the results highlight that explanation quality is influenced not only by the explainability method but also by the linguistic characteristics of the target domain.

This study is limited to base-sized transformer architectures, masking-based faithfulness evaluation, final-layer head-averaged attention extraction, and a fixed set of manually constructed bridge sentences. Although these design choices ensured methodological consistency

across all evaluated models, they may not fully represent explanation behaviour in larger language models or alternative evaluation settings. Future work should therefore extend the proposed framework to multilingual and multimodal transformer architectures, alternative attention extraction strategies, and causal or counterfactual explainability methods to further investigate explanation robustness under more diverse and realistic deployment scenarios.

Author Contributions: D.B.F. contributed to conceptualization, methodology, software development, data curation, formal analysis, investigation, visualization, and writing of the original draft. B.R.P.D. con-tributed to validation, supervision, manuscript review, and editing. All authors reviewed and approved the final version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The IMDb dataset is publicly available at https://huggingface.co/datasets/imdb [14]. The Hate Speech Offensive dataset is available at https://huggingface.co/datasets/hate_speech_offensive [13]. All experimental code will be made available at the corresponding author's repository upon acceptance.

Acknowledgments: The authors acknowledge the use of the HuggingFace Transformers library [5] and SHAP library [8] in supporting the experimental implementation and explainability analysis conducted in this study. The authors used AI-assisted language support tools during manuscript prep-aration for grammar refinement and language polishing. All technical content, experimental design, analysis, interpretation, and final manuscript validation were conducted solely by the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] J. Devlin, M.-W. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Oct. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[2] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, "Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1560–1569. doi: 10.18653/v1/2020.findings-emnlp.141.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *ArXiv*. pp. 2–6, Oct. 02, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[4] J. C. Timoneda and S. V. Vera, "BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text," *J. Polit.*, vol. 87, no. 1, pp. 347–364, Jan. 2025, doi: 10.1086/730737.

[5] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[6] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS One*, vol. 15, no. 8, p. e0237861, Aug. 2020, doi: 10.1371/journal.pone.0237861.

[7] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc Interpretability for Neural NLP: A Survey," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–42, Aug. 2023, doi: 10.1145/3546577.

[8] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230

[9] A. Vaswani *et al.*, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[10] S. Jain and B. C. Wallace, "Attention is not Explanation," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 3543–3556. doi: 10.18653/v1/N19-1357.

[11] S. Wiegreffe and Y. Pinter, "Attention is not not Explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 11–20. doi: 10.18653/v1/D19-1002.

[12] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A Survey of the State of Explainable AI for Natural Language Processing," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 447–459. doi: 10.18653/v1/2020.aacl-main.46.

[13] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 11, no. 1, pp. 512–515, May 2017, doi: 10.1609/icwsm.v11i1.14955.

[14] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2011, pp. 142–150. [Online]. Available: https://aclanthology.org/P11-1015/

[15] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *ArXiv*. Jan. 04, 2019. [Online]. Available: http://arxiv.org/abs/1711.05101
[16] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.
[17] N. A. Semary, W. Ahmed, K. Amin, P. Plawiak, and M. Hammad, "Improving sentiment classification using a RoBERTa-based hybrid model," *Front. Hum. Neurosci.*, vol. 17, Dec. 2023, doi: 10.3389/fnhum.2023.1292010.
[18] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 380–385. doi: 10.18653/v1/N19-1035.
[19] N. D. A. Saputra, M. Muljono, A. Karim, and D. R. I. M. Setiadi, "End-to-End Fine-Tuning of DeBERTa-Base for Stance Detection," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 698–715, Feb. 2026, doi: 10.62411/faith.3048-3719-168.
[20] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.
[21] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What Does BERT Look at? An Analysis of BERT's Attention," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 276–286. doi: 10.18653/v1/W19-4828.
[22] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4198–4205. doi: 10.18653/v1/2020.acl-main.386.
[23] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4902–4912. doi: 10.18653/v1/2020.acl-main.442.
[24] A. J. Böck, D. Slijepčević, and M. Zeppelzauer, "Exploring the Plausibility of Hate and Counter Speech Detectors with Explainable AI," in *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, Sep. 2024, pp. 1–8. doi: 10.1109/CBMI62980.2024.10859247.
[25] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "A Diagnostic Study of Explainability Techniques for Text Classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3256–3274. doi: 10.18653/v1/2020.emnlp-main.263.
[26] P. Hase and M. Bansal, "Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5540–5552. doi: 10.18653/v1/2020.acl-main.491.
[27] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 17, pp. 14867–14875, May 2021, doi: 10.1609/aaai.v35i17.17745.
[28] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1667–1682. doi: 10.18653/v1/2021.acl-long.132.