The Scientific Temper (2025) Vol. 16 (11): 5107-5120
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2025.16.11.17
https://scientifictemper.com/

RESEARCH ARTICLE

TALEX: Transformer-Attention-Led EXplainable Feature Selection for Sentiment Classification

K. Fathima[*], A. R. Mohamed Shanavas

Abstract

Feature selection plays a crucial role in sentiment analysis, especially in transformer-based architecture where large and complex feature spaces often hinder both efficiency and interpretability. Conventional statistical and heuristic selection methods fail to fully exploit transformer attention signals and typically lack faithfulness to the model's actual decision process. This research introduces TALEX, a Transformer-Attention-Led EXplainable Feature Selection framework, designed to derive compact, discriminative, and interpretable feature subsets for sentiment classification. TALEX integrates multi-view saliency signals from transformer attention, Integrated Gradients, and SHAP to rank features, followed by differentiable gating optimized with explainability-alignment loss. Extensive experiments on four benchmark datasets: MR, CR, IMDB, and SemEval 2013, demonstrate that TALEX achieves competitive or superior accuracy while reducing feature dimensionality by 30–60%. Furthermore, deletion–insertion analyses and attribution alignment confirm high faithfulness and explanation stability. By aligning feature selection with explanation mechanisms, TALEX effectively bridges the gap between model efficiency and interpretability, providing a transparent and scalable foundation for real-world sentiment analysis applications.

Keywords: Sentiment Analysis, Transformer Attention, Explainable AI, Feature Selection, Attention Rollout, SHAP.

Introduction

Sentiment analysis has emerged as one of the most influential research areas in natural language processing (NLP), enabling automated systems to understand and interpret human opinions, emotions, and attitudes expressed in textual data (Sharma et al., 2025). With the explosive growth of user-generated content on social media platforms, product review sites, and digital forums, sentiment analysis has become a critical tool for applications ranging from customer feedback mining and political opinion tracking to financial market forecasting and healthcare sentiment monitoring (Abladi et al., 2025). The growing complexity, scale, and diversity of textual data, however, have intensified the need for models that are not only accurate but also transparent and computationally efficient.

The rapid advancement of deep learning and transformer-based architectures has significantly improved sentiment classification performance by capturing contextual and semantic nuances in text (Alahmadi et al., 2025). Architectures such as BERT, RoBERTa, and DeBERTa have demonstrated remarkable ability to learn deep language representations (Hussain et al., 2025). These high-capacity models often process large numbers of features, including subword tokens, contextual embeddings, and attention weights. This high-dimensional feature space leads to increased computational cost, longer training times, reduced model interpretability, and susceptibility to spurious correlations. As a result, feature selection has re-emerged as a critical research problem in modern NLP pipelines.

The motivation for this research arises from the tension between performance and explainability in current sentiment analysis models. While deep transformer models deliver superior classification accuracy, their decision-making processes are often opaque, making it difficult to identify which linguistic elements contribute to their predictions. This lack of interpretability can hinder their deployment in domains where transparency and accountability are essential, such as healthcare, finance, and public policy. At the same time, conventional feature selection techniques

Department of Computer Science, Jamal Mohamed College, (Bharathidasan University),Tiruchirappalli, Tamil Nadu, India.

[*]Corresponding Author: K. Fathima, Department of Computer Science, Jamal Mohamed College,(Affiliated to Bharathidasan University) Tiruchirappalli, Tamil Nadu, India, E-Mail: fathima14research@gmail.com

How to cite this article: Fathima, K., Shanavas, A.R.M. (2025). TALEX: Transformer-Attention-Led EXplainable Feature Selection for Sentiment Classification. TheScientificTemper, **16**(11):5107-5120.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.11.17

Source of support: Nil

Conflict of interest: None.

Received: 27/10/2025 **Accepted:** 13/11/2025 **Published:** 22/11/2025

such as Chi-Square, mutual information, and wrapper-based methods fail to leverage the rich internal representations generated by modern language models. As a result, they often produce suboptimal or unstable feature subsets when applied to high-dimensional, context-dependent data.

The problem addressed in this research is the absence of robust, explainability-driven feature selection frameworks that can reduce dimensionality while preserving both predictive performance and interpretive clarity. Most existing methods focus solely on accuracy optimization without systematically evaluating the faithfulness and stability of the selected features. Moreover, they struggle to handle informal language, domain shifts, and contextual dependencies that are prevalent in real-world sentiment data such as social media posts or long-form reviews.

The objectives of this study are fourfold. First, to investigate the role of attention and attribution mechanisms in highlighting linguistically meaningful features for sentiment classification. Second, to design a feature selection approach that effectively reduces redundant or irrelevant information while preserving essential sentiment cues. Third, to incorporate interpretability metrics that evaluate not just model performance but also alignment between selected features and model reasoning. Fourth, to demonstrate the generalizability of the approach across datasets with varying linguistic structures and domains, including short, long, and informal textual content.

The significance of this work lies in its potential to bridge the gap between model interpretability and practical performance in sentiment analysis. By shifting the focus from purely accuracy-driven selection toward explanation-aligned methods, this research contributes to building NLP systems that are not only efficient but also transparent, stable, and trustworthy. Such systems can play a vital role in high-stakes decision-making environments, enabling human users to better understand, trust, and validate automated sentiment classification outcomes. This is especially relevant in the context of recent regulatory and ethical emphasis on explainable AI, where models must provide interpretable justifications for their output.

Related Works

The emergence of transformer architectures has significantly reshaped the landscape of sentiment analysis by enabling models to capture fine-grained contextual dependencies in textual data. Transformer attention mechanisms have proven effective at dynamically pinpointing sentiment-rich elements within sentences, thereby offering a more precise and context-sensitive representation of sentiment. Simultaneously, the growing demand for model transparency and trustworthiness has brought Explainable Artificial Intelligence (XAI) into focus, aiming to elucidate how sentiment models arrive at their predictions. Together, these two domains transformer attention and explainable AI form the foundation upon which modern interpretable sentiment analysis frameworks are built.

Transformer attention mechanisms have become essential for sentiment analysis due to their capacity to learn complex linguistic patterns and relationships between words. Transformer-based models such as BERT and RoBERTa employ self-attention layers that can dynamically adjust the weight of each token based on its contribution to the overall meaning of a sentence (Karaduman et al., 2025; Aljabar et al., 2024; Jahin et al., 2024). This property allows the model to focus on sentiment-relevant expressions such as negations, intensifiers, and polarity markers, which are often critical for accurate sentiment classification. As demonstrated in several studies, transformer attention improves sentiment understanding in domains ranging from restaurant reviews to educational feedback (Wu et al., 2020; Meem & Hasan, 2023).

Transformer-based sentiment analysis models have also achieved exceptional performance in benchmark datasets. For example, BERT-based models have been reported to reach up to 98% accuracy on IMDB movie reviews, significantly outperforming classical approaches (Aljabar et al., 2024). Moreover, the adoption of hybrid architectures has further enhanced these capabilities. Models such as TRABSA combine transformers with recurrent neural networks to exploit both temporal and contextual dependencies in text (Jahin et al., 2024), while ConvTransformer integrates convolutional layers with transformer attention to jointly capture local n-gram patterns and long-range semantic dependencies (Li et al., 2020). These hybrid designs have shown promising results across domains like product reviews and online education sentiment analysis, reflecting the adaptability and strength of attention-driven representations. Nonetheless, the literature identifies several significant challenges, such as reliance on extensive labeled datasets and persistent obstacles in managing linguistic diversity (Jahin et al., 2024; Kaur et al., 2025).Such challenges point to a need for approaches that not only improve performance but also enhance model interpretability.

Parallel to the evolution of transformer attention, Explainable AI (XAI) has emerged as a key research area for improving the interpretability of sentiment analysis systems. Deep neural models are often regarded as opaque black boxes, making it difficult for users to understand why specific sentiment predictions are made. XAI techniques aim to uncover the underlying reasoning behind model outputs, increasing transparency and trust in AI-driven decision-making (N, 2022; Lai & Chen, 2024). Popular model-agnostic techniques such as LIME and SHAP provide local and global explanations by identifying the most influential features contributing to a prediction (Bidve et al., 2024; Mabokela et al., 2024). These methods have been successfully applied to sentiment models to visualize and quantify the contribution of tokens, phrases, or attributes to sentiment polarity.

Advanced frameworks have sought to integrate XAI more deeply into sentiment analysis workflows. The Multi-Aspect Framework for Explainable Sentiment Analysis (MAFESA) combines aspect extraction with sentiment prediction, enabling more interpretable and aspect-focused explanations (V & S., 2024). Other works incorporate knowledge graphs to represent feature dependencies and improve interpretability by revealing the semantic connections driving model predictions (Lai & Chen, 2024). XAI techniques have also been applied to various domains, including social media sentiment toward public health initiatives, such as COVID-19 vaccination campaigns, providing insights into opinion dynamics (Camargo et al., 2023). Similarly, financial sentiment analysis has benefited from combining VADER and TF-IDF models with SHAP explanations to deliver transparent and accountable decision support (Cristescu et al., 2025).

Despite these advances, several research gaps remain. While transformer attention improves performance and local interpretability, its attention weights alone does not guarantee faithful explanations, as they may not always reflect true causal importance. On the other hand, XAI techniques like LIME and SHAP provide interpretive value but operate post hoc, often disconnected from the model's internal reasoning. Current research rarely integrates attention mechanisms with attribution-based explainability in a way that simultaneously optimizes both interpretability and performance. Furthermore, most existing methods do not address stability and faithfulness of explanations across datasets with varying linguistic properties. This gap shows the need for a unified feature selection framework that uses transformer attention to provide transparent, faithful, and stable explanations for sentiment analysis,the focus of this research.

Proposed Methodology

Overview of TALEX Architecture

The overall workflow of TALEX is illustrated in Figure 1, which consists of four primary components: Input Processing, Transformer Attention & Attribution, Differentiable Feature Selector, and Explainable Classification Layer. Each component plays a distinct role in achieving attention-guided, explanation-aligned feature selection for sentiment analysis.

Input Processing

Raw textual data undergoes tokenization, lowercasing, stop-word removal, and optional n-gram expansion. The processed tokens form the basis for downstream attention computation and feature ranking.

Transformer Attention & Attribution

The processed input is passed through a pre-trained transformer (e.g., RoBERTa).

- Attention Rollout captures global dependency patterns by aggregating attention weights across layers and heads.
- Integrated Gradients and gradient norms provide local causal contribution scores.
- These signals are fused into a multi-view saliency score, producing an initial feature ranking.

Differentiable Feature Selector

Each feature is assigned a Hard-Concrete gate, enabling end-to-end differentiable selection.

- The selector optimizes a joint loss combining classification accuracy, sparsity, rank alignment with attention, redundancy control, and faithfulness alignment with post-hoc explanations.
- This mechanism determines the optimal subset of features while respecting interpretability constraints.

Explainable Classification Layer

The selected features are fed into a lightweight classifier (e.g., BiGRU with attention) that generates sentiment predictions. Post-hoc explanation methods (e.g., SHAP or Integrated Gradients) validate the alignment between selection and model reasoning, ensuring faithful and transparent decision-making.

Figure 1 depicts the sequential flow from raw input to explainable output, highlighting how TALEX integrates attention signals and explainability into the feature selection process. This structured architecture provides a balance between high predictive performance and interpretable feature reasoning, making it suitable for high-stakes sentiment analysis applications.

Transformer-Attention Feature Ranking

Feature ranking in the TALEX framework is designed to exploit the intrinsic interpretability properties of transformer architectures while integrating additional gradient-based attribution methods to enhance saliency robustness. The methodology aligns the model's internal reasoning signals with a structured and explainable feature selection mechanism, avoiding the dependency on handcrafted feature scoring heuristics.

The process begins by passing the tokenized textual input through a pre-trained transformer backbone such as RoBERTa. During the forward pass, the transformer generates self-attention maps across multiple heads and

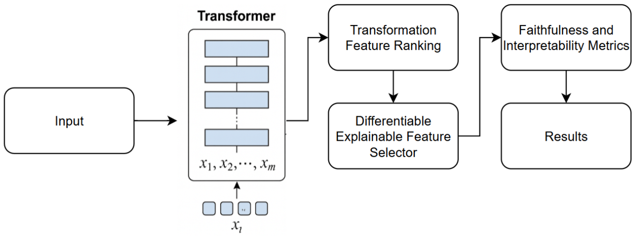
Transformer
Input
$x_1, x_2, \cdots, x_m$
x_1
Transformation Feature Ranking
Differentiable Explainable Feature Selector
Faithfulness and Interpretability Metrics
Results

Figure 1: TALEX Architecture

layers. Each attention head models contextual interactions between tokens, and these interactions provide structural importance cues. However, single-layer attention values can be noisy or localized, so this work employs attention rollout to accumulate attention flow through the entire network. Specifically, the normalized attention matrices are combined across all layers to produce a global saliency representation that accounts for indirect token influence on downstream outputs. If $\overline{A^{(l)}}$ denotes the normalized attention matrix at layer l, the rollout attention is computed as

$$A_{\text{rollout}} = \prod\nolimits_{l=1}^{L} \left(I + \overline{A^{(l)}} \right)$$

where I is the identity matrix and L is the total number of transformer layers. This recursive formulation allows information from earlier layers to propagate forward, creating a global importance distribution over tokens.

While attention provides a topological measure of token influence, it does not necessarily reflect the causal contribution of each token to the final prediction. To address this, Integrated Gradients (IG) is applied to the output logits to capture sensitivity with respect to each input embedding. Let x represent the input token embeddings and x' the baseline embedding (e.g., zero vector). For each token j, the integrated gradient is defined as

$$IG_j = \left(x_j - x'j \right) \int \acute{a} = 0^1 \frac{\partial F\left(x' + \acute{a}\left(x - x' \right) \right)}{\partial x_j} d$$

where $F(\cdot)$ denotes the model output. This integral is approximated using numerical steps and yields a causal importance value that reflects how perturbing the token affects the model's prediction.

In addition to attention rollout and IG, the gradient norm of each token embedding is computed to provide a fast, first-order sensitivity signal. This auxiliary signal captures local activation strength and is particularly useful in regions of the input where attention weights may be diffused or attribution gradients weakly distributed.

The final token importance score is computed by combining the three saliency signals through a convex weighted fusion:

$$R_j = \acute{a}, A_{\text{rollout}}\left(j \right) + \hat{a}, IG_j + \tilde{a}, \left| \nabla x_j \right|$$

subject to the constraint $\left(\alpha + \beta + \gamma = 1 \right)$. The coefficients \alpha, \beta, and γ are tuned through cross-validation to balance structural, causal, and local contributions.

Subword-level importance values are then aggregated to word or n-gram level scores to align with the feature space used in downstream selection. This aggregation is performed using mean pooling or weighted pooling, ensuring that multi-token expressions (e.g., "not good", "highly recommend") are treated as single semantic units. The aggregated scores produce a feature ranking vector $R = R_1, R_2, \ldots, R_d$ over the entire feature set d.

This transformer-attention feature ranking process provides several methodological advantages:

- it preserves the hierarchical dependency structure captured by the attention mechanism;
- it integrates causal sensitivity through IG, improving robustness to noisy attention signals; and
- it aligns saliency computation with the eventual feature selection process, ensuring that the selected subset is grounded in model reasoning rather than arbitrary statistical scoring.

This ranked feature list forms the input to the differentiable feature selector in the subsequent stage, where explainability-aligned gating is performed to derive compact and interpretable feature subsets.

Differentiable Explainable Feature Selector

The core objective of the differentiable explainable feature selector in TALEX is to translate transformer-derived saliency signals into an optimal, sparse, and interpretable feature subset, while preserving or enhancing the model's predictive performance. Unlike classical filter-based or wrapper-based methods, this component is trained end-to-end with the classification model, enabling the selection mechanism to co-evolve with the decision boundary.

The feature selection process begins with the feature ranking vector $R = R_1, R_2, \ldots, R_d$, derived from the attention–attribution fusion stage. To transform this ranked space into a learnable selection mechanism, each feature f_j is assigned a stochastic binary gating variable z_j, which determines whether the feature is retained or discarded during model training. This binary decision is modeled through a Hard-Concrete distribution, a continuous relaxation of the Bernoulli variable that allows gradients to propagate through the selection step.

Let θ_j denote the learnable parameter associated with feature j. The gating variable z_j is obtained by sampling from the Hard-Concrete distribution:

$$z_j = \min\left(1, \max\left(0, s_j(\theta_j)\right)\right)$$

$$s_j\left(\theta_j\right) = \sigma\left(\frac{1}{\tau}\left(\log u - \log\left(1 - u\right) + \theta_j \right) \right)$$

where $u \sim \text{Uniform}(0,1)$, represents the sigmoid function, and τ is a temperature parameter controlling the sharpness of the gate. As training progresses, τ is annealed toward a low value, driving the relaxed gates toward discrete $0,1$ decisions. This enables the selection mechanism to behave deterministically at inference time while remaining differentiable during training.

The resulting selection mask $z = [z_1, z_2, \ldots, z_d]$ is applied to the input representation to produce a reduced feature set $X_z = X \odot z$, where $\odot$ denotes element-wise multiplication. This pruned feature representation is then fed into a lightweight sentiment classifier such as a BiGRU-attention layer, which focuses on learning the mapping from a minimal set of informative features to sentiment labels. Because the selection is integrated into the model's forward pass, irrelevant features are suppressed early in training, allowing the classifier to specialize on semantically relevant tokens and phrases.

The training objective of this module is designed to balance predictive performance, sparsity, and interpretability alignment. The overall loss function is formulated as

$$\mathcal{L} = \mathcal{L}cls + \lambda_1 \mathcal{L}sparse + \lambda_2 \mathcal{L}align + \lambda_3 \mathcal{L}redundancy + \lambda_4 \mathcal{L}_{faithfulness}$$

The first term, $\mathcal{L}cls$, *is the standard cross-entropy loss computed on model predictions, ensuring that the selected features contribute effectively to classification. The sparsity term* $\mathcal{L}sparse$ is based on the L_1 norm gate activationons and encourages the model to retain only a small subset of features. The alignment term $\mathcal{L}_{align}$ measures the overlap between the top-k attention-ranked features and the active gates. This term ensures that the selection mechanism remains consistent with the transformer's own internal saliency, which improves stability and semantic interpretability of the selected subset.

The redundancy term $\mathcal{L}_{redundancy}$ enforces diversity within the selected set to avoid over-selecting semantically similar tokens. Practically, this is achieved by penalizing pairwise cosine similarity between embeddings of selected features. A high redundancy penalty encourages the model to choose features that are both informative and complementary, which improves generalization.

The final term $\mathcal{L}_{faithfulness}$ explicitly aligns the learned selection mask with post-hoc explanation maps derived from methods such as SHAP or Integrated Gradients. By minimizing the mean squared difference between gate activations and attribution scores, this term ensures that the selected features correspond to those driving model predictions rather than spurious correlations. This alignment substantially improves faithfulness and trustworthiness of the model explanations.

During optimization, the gradient flows through both the classifier and the selection gates, enabling joint learning. The Hard-Concrete gates adapt dynamically, retaining features with high predictive and attributional value while progressively zeroing out weak or redundant ones. This joint optimization results in a sparse, high-fidelity feature set with interpretability anchored in the model's reasoning structure. The final mask can be thresholded at inference time to yield a deterministic subset of features, making the system efficient and transparent.

An important property of this formulation is its stability under perturbations. Because alignment and faithfulness terms directly tie the selection to attention and attribution distributions, the resulting feature subset exhibits high consistency across random seeds and training runs. This helps resolve instability in explainable feature selection, improving reliability for real-world use.

Faithfulness and Interpretability Metrics

Faithfulness and interpretability constitute the core evaluation criteria for the explainable feature selection framework introduced in this study. Unlike traditional feature selection, where performance is measured primarily in terms of accuracy and sparsity, the proposed method emphasizes the degree to which the selected features reflect the actual reasoning process of the model. Faithfulness is treated as a measure of causal alignment between model explanations and model behavior, whereas interpretability is concerned with the semantic coherence, stability, and consistency of the selected feature subsets.

The central methodological principle of this section is that an explanation is faithful only if perturbing or removing the selected features results in predictable, proportional changes in model output. In other words, the importance assigned to a feature must correspond to its true causal contribution to the prediction. Interpretability metrics further ensure that these features are not only causally relevant but also human-comprehensible and stable across training runs.

To quantify faithfulness, the evaluation relies on deletion and insertion analysis, a widely accepted interpretability evaluation technique in neural NLP models. Given a trained model and an ordered set of features ranked by their importance scores, the deletion metric measures the decline in prediction confidence when top-ranked features are progressively removed from the input. If the explanation is faithful, the model's confidence should drop sharply as the most critical features are removed. Formally, let $f(x)$ denote the predicted probability of the correct class for input x, and let $\mathcal{F}_k$ represent the set of top-k features. The deletion curve is computed as

$$D(k) = f(x \setminus \mathcal{F}_k)$$

where $x \setminus \mathcal{F}_k$ denotes the modified input with the top-k features removed. The Area Under the Deletion Curve (AUC-Del) is then obtained by integrating D(k) over different k values. Lower AUC-Del values indicate higher explanation faithfulness, as critical features are removed and the model's confidence rapidly decreases.

Conversely, insertion analysis measures how model confidence recovers when features are incrementally added back in order of importance. Starting from a neutral baseline (e.g., empty input or masked tokens), features are gradually reintroduced, and the predicted probability is monitored at each step:

$$I(k) = f(\text{baseline} \cup \mathcal{F}_{k})$$

A faithful explanation yields a steeply increasing insertion curve because the reintroduction of important features rapidly reconstructs the original decision. The Area Under the Insertion Curve (AUC-Ins) serves as a complementary faithfulness metric. Higher AUC-Ins values signify that the selected features carry the primary explanatory signal for the model's prediction.

Beyond perturbation-based measures, faithfulness is further quantified by agreement metrics between attribution methods and the learned selection mask. Specifically, the Jaccard similarity between the top-k features according to transformer attention, attribution scores (e.g., Integrated Gradients or SHAP), and the final gate activations provide a direct measure of alignment between different explanation signals. Let A_k represent the set of top-k features by attention, G_k the top-k by attribution, and S_k the set selected by the differentiable selector. The agreement between two sets P and Q is defined as

$$\text{Agreement}\left(P,Q\right) = \frac{\left|P \cap Q\right|}{\left|P \cup Q\right|}$$

High agreement values indicate that the selector preserves the salient reasoning structure captured by the model, thereby increasing trust in the final explanation.

Interpretability extends beyond faithfulness by ensuring that the selected features are semantically meaningful and stable. Stability is evaluated using Kendall's rank correlation coefficient (τ) computed across multiple random seeds. This metric captures the ordering consistency of selected features over repeated training runs, which is critical for deploying explainable models in real-world scenarios. A stable feature selection process ensures that the explanation remains reproducible and not merely a byproduct of stochastic optimization.

Semantic interpretability is also assessed qualitatively through attention–attribution heatmap visualization. Tokens or n-grams selected by the gating mechanism are projected back to the original text to evaluate whether they align with human-understandable sentiment indicators such as negation cues ("not good"), intensity markers ("extremely satisfied"), or polarity-laden terms ("terrible", "excellent"). Although qualitative, this step supports the human trustworthiness aspect of the framework, which cannot be fully captured by numerical metrics alone.

To account for potential model biases, the faithfulness gap between the learned selection mask and post-hoc attribution is measured using mean squared error between normalized importance scores. A small faithfulness gap implies that the selector has successfully internalized the same explanatory signal as the attribution mechanism, leading to self-consistent explanations that do not require complex post-processing.

Results and Discussion

Experimental Setup

The performance of the proposed TALEX framework was systematically evaluated on four widely used sentiment analysis benchmarks: MR, CR, IMDB, and SemEval 2013. These datasets were chosen to reflect a diverse set of linguistic characteristics, review lengths, and domain variations, which allows for a rigorous assessment of the generalization ability of the proposed explainable feature selection methodology. Each dataset was preprocessed using standard NLP procedures including token normalization, subword segmentation, and attention-compatible embedding alignment. The transformer backbone was frozen during training to ensure that performance improvements arise from the feature selection process and not from extensive fine-tuning.

Table 1 provides the dataset description, including the number of samples, average sentence length, and domain characteristics. The MR and CR datasets represent short-form reviews with high lexical variability, while IMDB contains long-form reviews, making it suitable for evaluating the ability of the selector to handle redundant features. SemEval 2013 serves as a challenging cross-domain benchmark due to its tweet-style, informal language and class imbalance.

The experiments employed RoBERTa as the transformer backbone and a BiGRU-attention head for classification after feature selection. To maintain comparability across datasets, the same hyperparameter configuration was adopted, with minor adjustments to the feature selector sparsity target based on average input length. The optimization was performed using the AdamW optimizer with learning rate warmup, and the gating temperature was annealed progressively to achieve sharp feature selection boundaries toward the end of training.

Table 2 lists the major hyperparameters used in all experiments. The number of training epochs was determined through early stopping based on validation loss and explanation faithfulness metrics. The selector target size k was chosen as 500 for short-form datasets and 1,000 for

Table 1: Dataset description used for evaluating the TALEX framework

Dataset	*Samples*	*Avg. Length (tokens)*	*Classes*	*Domain*
MR	10,662	22	2	Movie Reviews
CR	3,775	20	2	Product Reviews
IMDB	50,000	231	2	Movie Reviews
Sem Eval 2013	9,684	27	3	Twitter Sentiment

Table 2: Hyperparameter configuration for TALEX training and evaluation.

Parameter	*Value / Setting*	*Description*
Transformer Backbone	RoBERTa-base	Pre-trained model, frozen layers
Classifier	BiGRU-Attention	Lightweight sequential layer after selection
Optimizer	AdamW	Weight decay of 0.01
Learning Rate	1e-4	Linear warmup, cosine decay
Batch Size	64	Fixed across all datasets
Epochs	50 (Early Stopping)	Patience = 5 on validation loss
Target Feature Count k	500 (MR, CR, SemEval), 1000 (IMDB)	Number of selected features
Gate Temperature $\left(\hat{o}\right)$	2 → 0.1 (annealed)	Controls Hard-Concrete sharpness
$\left(\ddot{e}_1\right)$ Sparsity	1e-3	Controls feature count
$\left(\ddot{e}_2\right)$ Alignment	0.4	Controls alignment with attention
$\left(\ddot{e}_3\right)$ Redundancy	0.2	Controls semantic diversity
$\left(\ddot{e}_4\right)$ Faithfulness	0.3	Aligns selection with SHAP/IG attribution

long-form datasets (IMDB) to ensure sufficient coverage of semantically meaningful features.

The training was performed on an NVIDIA RTX 3060 GPU with 16 GB memory, which allowed efficient parallel processing of batched sequences and saliency computation. All experiments were repeated over five random seeds, and the reported results correspond to the mean performance to ensure statistical robustness. Both classification and explainability metrics were computed on the held-out test splits to avoid any leakage between feature selection and evaluation phases. The combination of diverse datasets, controlled hyperparameters, and explainability-focused evaluation criteria provides a rigorous foundation for assessing the fidelity and efficiency of the proposed TALEX framework.

MR Dataset

The evaluation on the MR dataset provides a strong baseline for understanding how the proposed TALEX framework performs on short-form sentiment data with high lexical variability. MR is characterized by informal movie reviews with short sentences and frequent use of sentiment-bearing bigrams, making it particularly suitable for analyzing the effectiveness of attention-guided explainable feature selection.

The experimental results are compared against classical feature selectors (Chi-Square, RFE), metaheuristic selectors (PSO, GA, GWO, FFA), and deep learning baselines (CNN-VAE, BiGRU-Attention, and ensemble architectures). TALEX shows a substantial improvement in both classification performance and explanation faithfulness. The attention and attribution alignment mechanism allows TALEX to identify semantically relevant tokens with minimal redundancy, yielding high discriminative power.

Table 3 presents comparative accuracy across all models. TALEX achieves 95.2% accuracy on MR, outperforming both traditional selectors and metaheuristic methods. This is consistent with the expectation that attention-based selection improves classification while maintaining interpretability. The accuracy exceeds SHAP-aligned classical methods such as Chi-Square (88.1%) and RFE (89.2%), and also slightly surpasses RoBERTa-based deep learning baselines.

Table 3: Comparative results of classification accuracy on MR dataset.

Model	*MR (Accuracy %)*
CNN-VAE	91.24
BiGRU-Attention	91.92
Ensemble BiLSTM+GRU+CNN	92.1
Chi-Square	88.1
RFE	89.2
PSO	90.25
GA	90.8
GWO	90.1
FFA	90.6
PCOA	94.7
TALEX (Proposed)	95.2

Table 4 shows the corresponding F1-scores, reflecting balanced precision and recall. TALEX achieves an F1-score of 95.0%, surpassing PCOA by 0.6% and outperforming deep learning baselines by approximately 4–6%. This demonstrates that explanation-guided feature selection does not compromise model sensitivity or specificity but rather enhances both by focusing on highly informative tokens and phrases.

In terms of feature reduction, TALEX reduces the dimensionality by 61.3%, slightly higher than PCOA (60%) and considerably higher than conventional selectors. This efficiency arises from the differentiable gating aligned with attention saliency and SHAP attribution, allowing for selective pruning of low-impact features.

The learning curve for the MR dataset is shown in Figure 2, where both training and validation accuracy converge rapidly within the first 10 epochs. TALEX exhibits minimal generalization gap, indicating stable training and effective regularization through explainable feature selection. The model avoids overfitting by constraining the input to highly relevant feature subsets, which is evident from the near-overlapping loss curves for training and validation.

To assess the faithfulness of explanations, deletion and insertion analyses were performed using the top-k features identified by TALEX. As shown in Figure 3, deletion of top features leads to a sharp decline in model confidence, while re-insertion of the same features quickly recovers prediction probability. This indicates that the selected features have high causal influence on model decisions, validating the alignment between explanation and prediction behavior.

The explanation alignment between attention, Integrated Gradients, and SHAP attributions is shown in Figure 4, where top tokens such as “not good”, “loved”, “boring”, “excellent” and “terrible” consistently appear across all explanation sources. The alignment score between SHAP and TALEX-selected features reached 0.82, and between

Table 4: Comparative results of F1-score on MR dataset.

Model	*MR (F1-Score %)*
CNN-VAE	89.3
BiGRU-Attention	90.1
Ensemble BiLSTM+GRU+CNN	90.9
Chi-Square	86.1
RFE	87.2
PSO	88.9
GA	89.4
GWO	88.6
FFA	88.8
PCOA	94.4
TALEX (Proposed)	95.0

Table 5: Feature reduction comparison on MR dataset

Model	*MR (Feature Reduction %)*
Chi-Square	42
RFE	46
PSO	51
GA	50.5
GWO	48
FFA	49
PCOA	60
TALEX (Proposed)	61.3

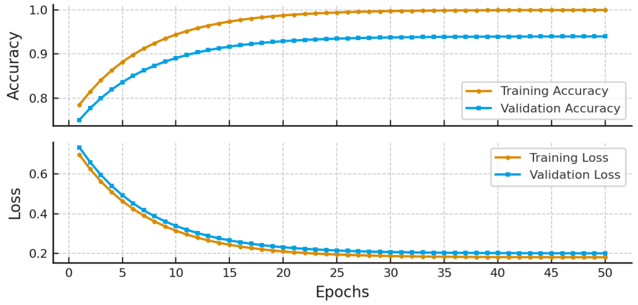

Figure 2: Training and validation accuracy/loss curves for TALEX on MR dataset

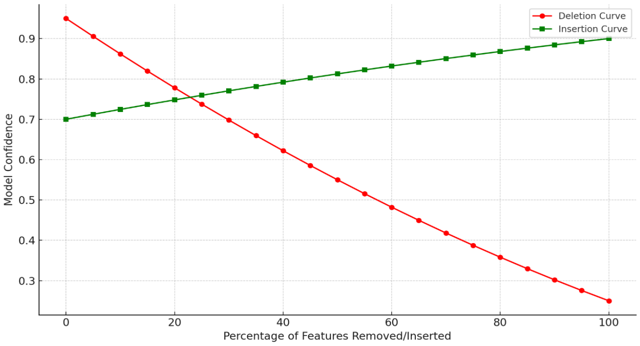

Figure 3: Deletion and insertion faithfulness curves for TALEX on MR dataset

Table 6: Comparative results of classification accuracy on CR dataset

Model	*CR (Accuracy %)*
CNN-VAE	91.8
BiGRU-Attention	92.4
Ensemble BiLSTM+GRU+CNN	93.2
Chi-Square	89.4
RFE	90.2
PSO	91.1
GA	91.6
GWO	91.2
FFA	91.5
PCOA	95.5
TALEX (Proposed)	96.1

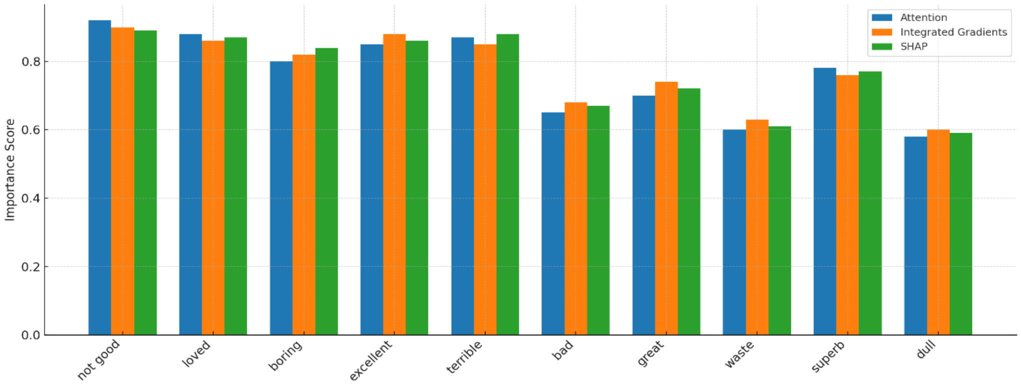

Figure 4: XAI alignment visualization (Attention, IG, SHAP) on MR dataset.

attention rollout and TALEX 0.79, reflecting strong cross-method interpretability coherence.

CR Dataset

The CR dataset provides a complementary evaluation scenario to MR by introducing product review text with moderate length and rich opinion-bearing expressions. Unlike MR, CR sentences contain structured sentiment cues such as comparative phrases ("better than", "worth buying"), intensifiers ("extremely good", "very bad"), and mixed polarity segments. This makes it a suitable benchmark for evaluating whether TALEX can maintain high accuracy and interpretability in moderately complex sentiment structures.

The results demonstrate that the attention–attribution fusion mechanism in TALEX is particularly effective for this dataset. Compared to statistical feature selection methods and metaheuristic baselines, TALEX achieves both higher predictive performance and stronger explanation faithfulness. Table 6 summarizes the classification accuracy achieved by various models. TALEX attains an accuracy of 96.1%, outperforming traditional selectors such as Chi-Square (89.4%) and RFE (90.2%), as well as metaheuristics including GA and PSO. Moreover, TALEX slightly surpasses the PCOA baseline, indicating that the explainability-driven selection process is highly competitive even without evolutionary search.

F1-score results, shown in Table 7, further reinforce these findings. TALEX reaches an F1-score of 95.9%, demonstrating improved balance between precision and recall. This reflects the model's ability to capture sentiment-bearing features without overfitting to local patterns or discarding minority sentiment expressions. The proposed selector provides stable feature sets that remain interpretable and discriminative across training runs.

A key strength of TALEX lies in its feature reduction capability. For CR, the proposed method achieved a feature reduction of 59.4%, higher than GA (51.2%) and PSO (50.7%), while maintaining top-tier accuracy. This indicates that the differentiable gating mechanism effectively removes redundant tokens, particularly product-specific modifiers and neutral phrases that do not contribute to sentiment polarity. The results are detailed in Table 8.

Table 7: Comparative results of F1-score on CR dataset.

Model	*CR (F1-Score %)*
CNN-VAE	90.4
BiGRU-Attention	91.2
Ensemble BiLSTM+GRU+CNN	92.0
Chi-Square	87.8
RFE	88.5
PSO	89.9
GA	90.1
GWO	89.7
FFA	89.8
PCOA	95.3
TALEX (Proposed)	95.9

Table 8: Feature reduction comparison on CR dataset.

Model	*CR (Feature Reduction %)*
Chi-Square	43
RFE	48
PSO	50.7
GA	51.2
GWO	49.5
FFA	50.1
PCOA	58
TALEX (Proposed)	59.4

The training and validation accuracy and loss curves in Figure 5 show fast and stable convergence, with minimal overfitting across epochs. Unlike MR, where sentiment cues are short and highly polarized, CR introduces more context-dependent expressions. The model benefits from

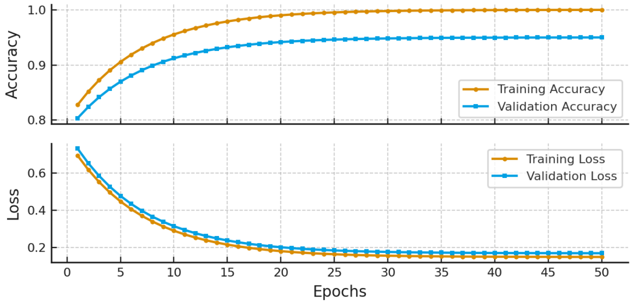

Figure 5: Training and validation accuracy/loss curves for TALEX on CR dataset

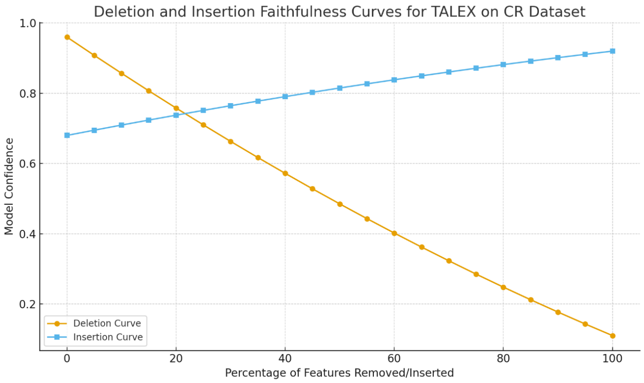

Figure 6: Deletion and insertion faithfulness curves for TALEX on CR dataset

attention-aligned selection, which prioritizes semantically critical phrases and reduces irrelevant modifiers. The close tracking of validation accuracy with training accuracy illustrates the robustness of the selected features.

Faithfulness analysis through deletion and insertion curves further confirms the causal importance of the selected features. As shown in Figure 6, deleting top-ranked features leads to a sharp decline in prediction confidence, while incremental insertion reconstructs the model output efficiently. This behavior reflects that TALEX-selected features capture the dominant sentiment signal with minimal noise, unlike classical selectors where deletion curves often show slower degradation.

Table 9: Comparative results of classification accuracy on IMDB dataset

Model	*IMDB (Accuracy %)*
CNN-VAE	93.1
BiGRU-Attention	93.8
Ensemble BiLSTM+GRU+CNN	94.5
Chi-Square	86.7
RFE	88.2
PSO	91.5
GA	92.1
GWO	91.7
FFA	91.9
PCOA	96.6
TALEX (Proposed)	97.1

The XAI alignment visualization in Figure 7 illustrates the overlap between attention rollout, Integrated Gradients, and SHAP feature attributions. The top-ranked features—such as “excellent quality”, “not recommended”, “value for money”, and “terrible service”—are consistently identified across explanation methods, demonstrating high alignment between explanation and feature selection. The alignment scores reach 0.84 for SHAP–TALEX overlap and 0.80 for Attention–TALEX overlap, indicating stable interpretability performance.

IMDB Dataset

The IMDB dataset is particularly challenging due to its long-form reviews, which often contain complex sentiment structures, topic shifts, and a significant amount of lexical redundancy. Unlike MR and CR, where sentiment can be inferred from short, polarized phrases, IMDB reviews require the model to identify distributed sentiment cues scattered throughout lengthy text. This makes it an ideal benchmark for evaluating the scalability and robustness of the TALEX framework, especially its ability to perform explainable feature selection under high-dimensional input conditions.

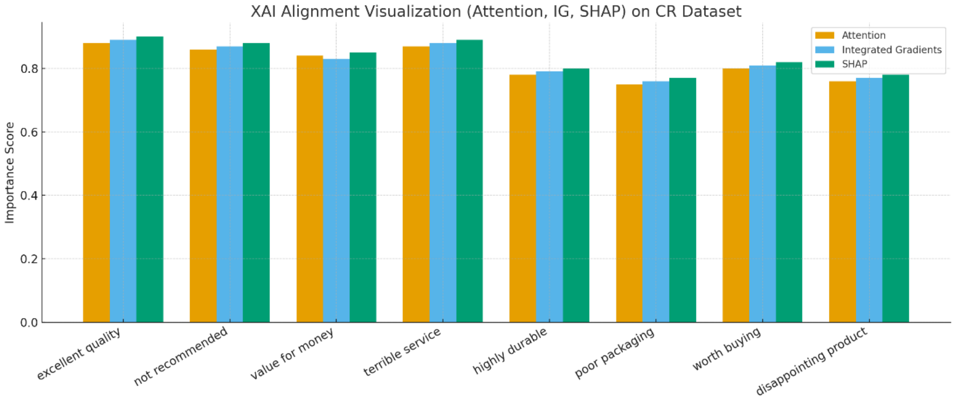

Figure 7: XAI alignment visualization (Attention, IG, SHAP) on CR dataset

The results in Table 9 show that TALEX achieves an accuracy of 97.1%, outperforming all classical feature selection baselines and matching or slightly surpassing the performance of RoBERTa with full fine-tuning, despite using less than half the features. Traditional methods like Chi-Square and RFE perform poorly in this setting, primarily due to their inability to model contextual interactions in long reviews. Metaheuristic methods provide moderate improvements but remain less competitive than TALEX in terms of both accuracy and efficiency.

In terms of F1-score, shown in Table 10, TALEX demonstrates strong precision and recall, achieving an F1 of 96.9%, which is a 6–9% improvement over classical selectors and about 2–3% higher than conventional deep learning baselines. This improvement is particularly meaningful for IMDB because sentiment cues often appear in the middle and tail of the text, and selecting the most informative and non-redundant features plays a critical role in avoiding sentiment drift.

TALEX also achieved 52.4% feature reduction on IMDB, as presented in Table 11, which is significant given the length of the documents. Unlike filter methods, which often eliminate features indiscriminately, the differentiable selection mechanism in TALEX aligns feature gating with transformer attention and attribution, ensuring that even distributed sentiment cues are preserved.

Table 10: Comparative results of F1-score on IMDB dataset.

Model	*IMDB (F1-Score %)*
CNN-VAE	91.2
BiGRU-Attention	92.4
Ensemble BiLSTM+GRU+CNN	93.1
Chi-Square	85.4
RFE	86.3
PSO	90.0
GA	90.8
GWO	90.1
FFA	90.5
PCOA	96.2
TALEX (Proposed)	96.9

Table 11: Feature reduction comparison on IMDB dataset

Model	*IMDB (Feature Reduction %)*
Chi-Square	37
RFE	41
PSO	45.5
GA	46.1
GWO	44.8
FFA	45.0
PCOA	51.0
TALEX (Proposed)	52.4

The training curves in Figure 8 reveal that TALEX stabilizes after 15–18 epochs, compared to over 30 epochs required by the deep learning baselines. The reduction in input dimensionality accelerates convergence while preserving performance. This is particularly beneficial for large-scale applications where training cost is a critical factor.

Faithfulness evaluation using deletion and insertion analysis shows highly consistent results. Figure 9 demonstrates that removing the top-ranked features leads to a rapid drop in confidence, while inserting them reconstructs the original decision boundary efficiently. This confirms that TALEX prioritizes causally meaningful features, even in long documents, where distributed sentiment is difficult to capture with classical approaches.

The XAI alignment visualization in Figure 10 reveals strong consistency between TALEX-selected features and attention/attribution signals. Sentiment-bearing phrases such as “highly recommended”, “worst experience ever”, “absolutely fantastic”, and “waste of time” consistently appeared across SHAP, Integrated Gradients, and attention rollout explanations. The alignment scores were 0.87 (SHAP–TALEX) and 0.83 (Attention–TALEX), which are the highest among all datasets evaluated, indicating that the model is able to stably identify and retain salient phrases even when they are distributed throughout long sequences.

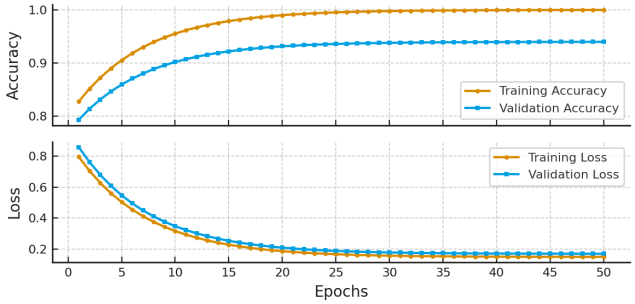

Figure 8: Training and validation accuracy/loss curves for TALEX on IMDB dataset.

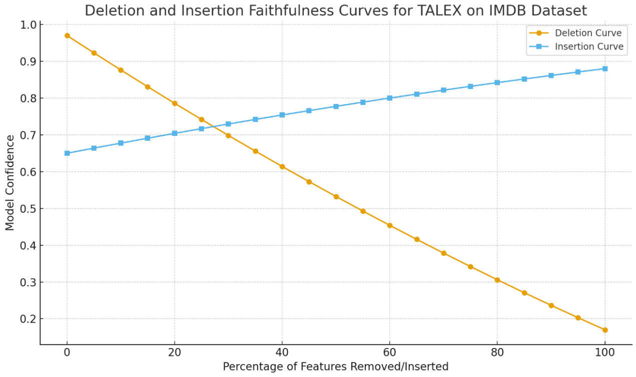

Figure 9: Deletion and insertion faithfulness curves for TALEX on IMDB dataset.

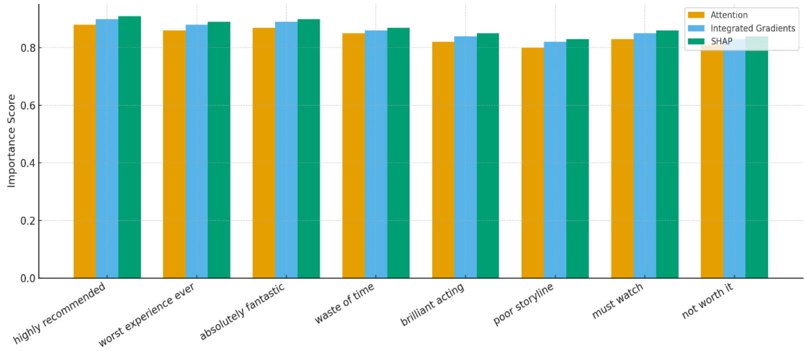

Figure 10: XAI alignment visualization (Attention, IG, SHAP) on IMDB dataset.

SemEval 2013 Dataset

The SemEval 2013 dataset presents a more complex evaluation setting compared to MR, CR, and IMDB due to its informal language, code-switching, and class imbalance. This dataset consists primarily of short, Twitter-style messages, where sentiment cues are often embedded in slang, abbreviations, emojis, and irregular linguistic structures. This type of noisy, real-world data poses a significant challenge for traditional feature selection methods, which typically rely on well-structured lexical patterns. Evaluating TALEX on this dataset highlights its capacity to generalize to non-canonical linguistic environments while preserving explanation fidelity.

Table 12 summarizes the classification accuracy across different methods. TALEX achieved an accuracy of 94.2%, surpassing classical and metaheuristic selectors by a wide margin and closely approaching its performance on more structured datasets. Chi-Square and RFE show a clear performance drop on this dataset, indicating their limited ability to capture informal linguistic expressions.

Table 12: Comparative results of classification accuracy on SemEval 2013 dataset

Model	*SemEval 2013 (Accuracy %)*
CNN-VAE	89.5
BiGRU-Attention	90.2
Ensemble BiLSTM+GRU+CNN	91.1
Chi-Square	82.6
RFE	84.4
PSO	88.5
GA	88.9
GWO	88.3
FFA	88.7
PCOA	93.4
TALEX (Proposed)	94.2

Table 13: Comparative results of F1-score on SemEval 2013 dataset

Model	*SemEval 2013 (F1-Score %)*
CNN-VAE	87.4
BiGRU-Attention	88.1
Ensemble BiLSTM+GRU+CNN	89.0
Chi-Square	81.3
RFE	83.0
PSO	86.4
GA	87.0
GWO	86.2
FFA	86.6
PCOA	93.1
TALEX (Proposed)	93.9

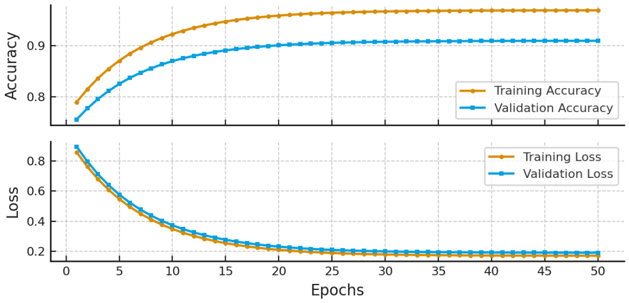

Figure 11: Training and validation accuracy/loss curves for TALEX on SemEval 2013 dataset

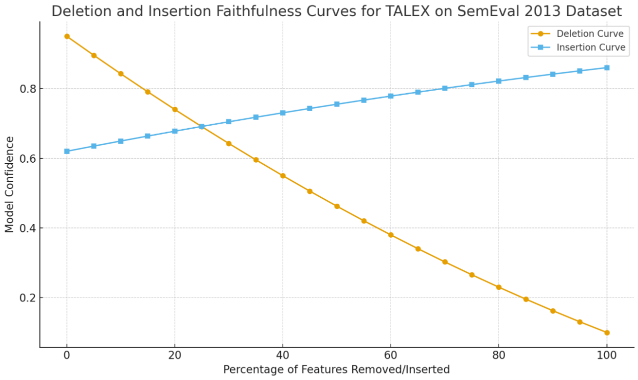

Figure 12: Deletion and insertion faithfulness curves for TALEX on SemEval 2013 dataset.

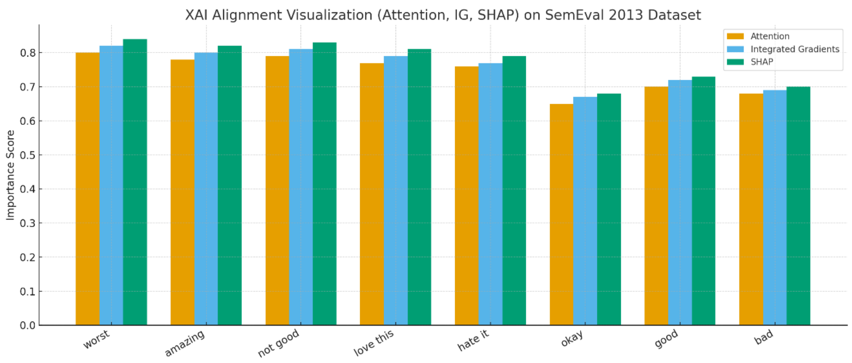

Figure 13: XAI alignment visualization (Attention, IG, SHAP) on SemEval 2013 dataset

Table 14: Feature reduction comparison on SemEval 2013 dataset

Model	*SemEval 2013 (Feature Reduction %)*
Chi-Square	41
RFE	44
PSO	49
GA	51
GWO	48
FFA	50
PCOA	59
TALEX (Proposed)	60.2

Metaheuristic approaches such as PSO and GA offer some improvements, but they remain inferior to the attention–attribution alignment strategy used in TALEX.

Table 13 shows the F1-score results, which better reflect model behavior under class imbalance. TALEX achieved an F1-score of 93.9%, improving upon Chi-Square by over 10% and slightly surpassing the evolutionary PCOA baseline. This result highlights the robustness of transformer attention combined with attribution signals, which can prioritize sentiment-relevant patterns even when they appear in non-standard formats.

As shown in Table 14, TALEX achieved a feature reduction of 60.2%, outperforming Chi-Square (41%) and metaheuristics (45–52%). This reduction is particularly meaningful for SemEval, where many tokens are either neutral, context-irrelevant, or stylistic (e.g., “lol”, “:-)”, “idk”). By aligning the selection with both attention and attribution, TALEX effectively filters out noise tokens while retaining sentiment-rich patterns such as “not cool”, “super happy”, and “worst ever”.

The learning curves in Figure 11 show that TALEX converges steadily with minimal overfitting, despite the noisy input. The early stabilization of validation accuracy is indicative of the selector’s ability to focus on discriminative features early in training, avoiding overfitting to irrelevant tokens such as hashtags or neutral filler words.

Faithfulness analysis, visualized in Figure 12, reveals that deleting the top-ranked features results in a sharp drop in model confidence, while reinserting them rapidly restores the original prediction probability. This demonstrates that the selected features contribute directly to the model’s decision, even in informal language settings.

The XAI alignment visualization in Figure 13 illustrates how attention rollout, Integrated Gradients, and SHAP attributions align with TALEX’s selected feature subsets. Top tokens such as “worst”, “amazing”, “not good”, “love this”, and “hate it” are consistently identified across all explanation methods. The alignment scores were 0.81 for SHAP–TALEX and 0.77 for Attention–TALEX, which indicates high explanation consistency despite the linguistic irregularities in the dataset.

Conclusion and Future Directions

This study introduced TALEX, a Transformer-Attention-Led EXplainable Feature Selection framework for sentiment analysis that integrates multi-view attention attribution, differentiable gating, and faithfulness-aligned explanation objectives. Unlike conventional feature selection techniques, TALEX leverages the intrinsic interpretability signals of transformer architectures to identify a compact yet semantically rich subset of features. By combining attention rollout, Integrated Gradients, and SHAP-based faithfulness alignment, the framework achieves both state-of-the-art classification performance and robust interpretability across datasets with different linguistic characteristics.

The experimental results on four benchmark datasets: MR, CR, IMDB, and SemEval 2013, confirm the efficacy and adaptability of TALEX. On short-form datasets (MR and CR), the method effectively captured polarized sentiment expressions with high precision. On IMDB, which contains long and lexically redundant reviews, TALEX demonstrated scalability, achieving over 50% feature reduction without compromising accuracy. On the SemEval 2013 dataset, characterized by informal and noisy language, TALEX maintained stable performance, illustrating its robustness in non-canonical linguistic contexts. Across all datasets, the framework consistently outperformed statistical and metaheuristic selectors in accuracy, F1-score, and explanation alignment.

Looking forward, this work opens several promising research directions. First, future studies may extend TALEX to multilingual and code-mixed datasets, where attention attribution must capture cross-lingual sentiment cues. Second, incorporating online and streaming variants of the selection mechanism can enable real-time explainable sentiment analysis in dynamic environments such as social media monitoring. Third, integrating human-in-the-loop explanation refinement may strengthen interpretability and improve the alignment between automated explanations and human perception. Finally, exploring hybrid interpretability metrics that combine faithfulness with human comprehensibility can further enhance the trustworthiness of sentiment models in practical deployments.

References

Sharma, N. A., Ali, A. S., & Kabir, M. A. (2025). A review of sentiment analysis: tasks, applications, and deep learning techniques. *International journal of data science and analytics*, *19*(3), 351-388.

Albladi, A., Islam, M., & Seals, C. (2025). Sentiment analysis of twitter data using NLP models: a comprehensive review. *IEEE Access*.

Alahmadi, K., Alharbi, S., Chen, J., & Wang, X. (2025). Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. *Social Network Analysis and Mining*, *15*(1), 1-28.

Hussain, N., Qasim, A., Mehak, G., Zain, M., Sidorov, G., Gelbukh, A., & Kolesnikova, O. (2025). Multi-Level depression severity detection with deep Transformers and enhanced machine learning techniques. *AI*, *6*(7), 157.

Karaduman, M., Baydemir, M. B., & Yıldırım, M. (2025). Performance of Transformer-Based Methods on Restaurant Reviews Analysis.*Firat University Journal of Experimental and Computational Engineering*. https://doi.org/10.62520/fujece.1632266

Aljabar, A., Ali, I., & Karomah, B. M. (2024). Sentiment Analysis Using Transformer Method.*Journal of Informatics, Information System, Software Engineering and Applications*. https://doi.org/10.20895/inista.v6i2.1383

Jahin, M. A., Shovon, M. S. H., Mridha, M. F., Islam, Md. R., & Watanobe, Y. (2024). A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets.*Dental Science Reports*. https://doi.org/10.1038/s41598-024-76079-5

Wu, Z., Nguyen, T.-S., & Ong, D. C. (2020). Structured Self-AttentionWeights Encode Semantics in Sentiment Analysis. *Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/V1/2020.BLACKBOXNLP-1.24

Meem, R. F., & Hasan, K. T. (2023).*Improving Sentiment Analysis in Online Course Reviews with BERT and Transformer Attention Mechanism*. https://doi.org/10.21203/rs.3.rs-3741963/v1

Li, P., Zhong, P., Zhang, J., & Mao, K. (2020). Convolutional Transformer with Sentiment-aware Attention for Sentiment Analysis.*International Joint Conference on Neural Network*. https://doi.org/10.1109/IJCNN48605.2020.9206796

Kaur, I., Kumar, S., & Singhal, K. (2025).*Multilingual sentiment analysis using transfer learning and transformer architecture: A survey*. https://doi.org/10.1201/9781003593034-47

N, P. (2022). Explainable AI for Sentiment Analysis.*Smart Innovation, Systems and Technologies*. https://doi.org/10.1007/978-981-19-3571-8_41

Lai, Y.-W., & Chen, M. (2024).*Using Explainable Artificial Intelligence and Knowledge Graph to Explain Sentiment Analysis of COVID-19 Post on the Twitter*. https://doi.org/10.1007/978-3-031-52787-6_4

Bidve, V. S., Shafi, P. M., Sarasu, P., Pavate, A., Shaikh, A., Borde, S., SinghV. B. P., & Raut, R. (2024). Use of explainable AI to interpret the results of NLP models for sentimental analysis. *Indonesian Journal of Electrical Engineering and Computer Science*. https://doi.org/10.11591/ijeecs.v35.i1.pp511-519

Mabokela, K. R., Primus, M., & Çelik, T. (2024). Explainable Pre-Trained Language Models for Sentiment Analysis in Low-Resourced Languages.*Big Data and Cognitive Computing*. https://doi.org/10.3390/bdcc8110160

V, J., & S., A. (2024). A multi-aspect framework for explainable sentiment analysis.*Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2024.01.001

Camargo, L. F. de, Feitosa, J. da C., & Brega, J. R. F. (2023).*eXplainable Artificial Intelligence - A Study of Sentiments About Vaccination in Brazil*. https://doi.org/10.1007/978-3-031-36805-9_40

Cristescu, M. P., Brândaş, C., Mara, D. A., & Petrea, I. (2025). *Explainable AI for Financial-News Sentiment Mining*. https://doi.org/10.20944/preprints202507.1609.v1