Integrating Attention Attribution and Pretrained Language Models for Transparent Discriminative Learning

Xiangchen Song *

Posted Date: 5 February 2026

doi: 10.20944/preprints202602.0344.v1

Keywords: discriminative learning; attention attribution; interpretability; sensitivity analysis

 doi:10.20944/preprints202602.0344.v1

Article

Integrating Attention Attribution and Pretrained Language Models for Transparent Discriminative Learning

Xiangchen Song

University of Michigan, Ann Arbor, USA; xiangchen.sxc@gmail.com

Abstract

This paper addresses the problem of insufficient interpretability in discriminative learning and proposes an interpretable discriminative learning method based on attention attribution. The study builds on the representation power of pretrained language models and introduces a multi-head attention mechanism to capture both global and local semantic dependencies, thereby obtaining richer feature representations. On this basis, attention attribution is used to calculate the importance scores of input features during prediction, and the attribution distribution reveals the core semantic cues relied upon by the model in discrimination, which enhances the transparency of the decision process. The framework consists of input embedding, contextual modeling, attribution feature extraction, and a classification layer, enabling the model to provide clear explanatory paths while maintaining high discriminative accuracy. The method is systematically validated under multiple single-factor sensitivity settings, including the effects of learning rate, sentence order disruption, dropout rate, and class imbalance on model performance and robustness. Experimental results show that the method achieves stability and superiority in metrics such as accuracy, precision, recall, and F1-score, maintaining strong discriminative ability and consistent interpretability under different conditions. Overall, by integrating attention mechanisms with attribution methods, this paper achieves a balance between performance and interpretability and provides an effective solution for discriminative learning in complex contexts.

Keywords: discriminative learning; attention attribution; interpretability; sensitivity analysis

1. Introduction

In the development of natural language processing, text understanding and discrimination have always been central issues [1]. With the rise of deep learning and pretrained language models, the ability to capture contextual dependencies and semantic details from large-scale corpora has been greatly enhanced. This provides a solid foundation for tasks such as classification, retrieval, and generation. However, despite remarkable progress in performance, the internal decision-making mechanisms of these models often remain a "black box." It is difficult to explain how the model makes judgments in complex semantic spaces. This limitation not only affects the credibility of models in critical applications but also undermines their sustainable value in both research and practice. Enhancing interpretability while maintaining discrimination performance has therefore become an important research direction [2].

In the context of discriminative learning, models must not only distinguish between categories but also remain stable and robust in scenarios where semantic boundaries are blurred and differences are subtle. Such challenges are common in areas like financial opinion monitoring, medical text analysis, and policy document interpretation. In these cases, the transparency of model decisions directly affects the acceptability and compliance of results. If the basis of discrimination cannot be made explicit, then even accurate predictions may be rejected due to a lack of verifiability. Integrating

 doi:10.20944/preprints202602.0344.v1

interpretability into discriminative learning is thus not only a technical pursuit but also a practical requirement for the healthy deployment of artificial intelligence in society [3].

Recently, attention mechanisms, as core components of pretrained models, have shown natural potential for interpretability. They assign weights to input features and provide semantic attribution paths. Yet a single attention distribution has clear limitations [4]. It may be influenced by hierarchical structures, sparse context, or redundant information, and may fail to fully reflect the true reasoning behind discrimination. To address this, combining attention attribution with contrastive signals has become a new research direction. Contrastive signals highlight differences in attention patterns between positive and negative samples, aligning explanations more closely with actual decision-making. This approach not only increases the credibility of interpretations but also opens new paths for human–machine interaction and system optimization.

From the perspective of applications, interpretable discriminative learning is of particular importance. In financial risk detection, regulators must understand which textual features led to a risk judgment to validate it with expert knowledge. In medical scenarios, doctors need to know the semantic cues the model focused on to support diagnosis and treatment decisions. In legal and policy analysis, transparent model explanations enhance credibility and reduce risks of algorithmic bias or discrimination. Research combining attention attribution with contrastive signals not only advances theoretical understanding of interpretability but also provides practical safeguards for compliance and trust [5,6].

Recent studies on multimodal information integration further indicate that jointly modeling heterogeneous signals within a unified representation space can significantly improve prediction stability and structural consistency [7]. This finding suggests that transparent modeling mechanisms are particularly important when dealing with complex and multi-source data.

In conclusion, interpretable discriminative learning based on attention attribution and contrastive signals is both a necessary trend in natural language processing and a key step toward real-world deployment of artificial intelligence. It helps move models from "black box" to "transparent," responding to urgent demands for interpretability and trust. At the same time, it provides strong technical support for interdisciplinary applications. By pursuing this direction, researchers can narrow the gap between performance and interpretability, allowing language models to realize their full value in broader contexts and driving intelligent systems toward understandability, reliability, and sustainability [8].

2. Related Work

In natural language processing, research on discriminative models has long focused on improving classification and recognition performance [9]. Early methods relied on traditional machine learning and handcrafted features. They used statistical features, bag-of-words models, or syntactic structures to achieve discrimination. However, these methods often failed to capture complex contextual dependencies [10]. As a result, they performed poorly when dealing with subtle semantic differences or domain-specific expressions. With the development of deep learning, neural network-based models have gradually replaced traditional methods. They showed a stronger ability to capture semantic hierarchies and contextual information. Recurrent networks and convolutional structures in particular provided new possibilities for sequence modeling and local pattern extraction. Yet these methods still faced problems of unclear decision boundaries and limited interpretability, which could not meet the demand for transparency and stability in highly sensitive applications [11].

The emergence of pretrained language models has greatly advanced discriminative learning. By training on large-scale corpora in an unsupervised manner, these models can generate highly semantic contextual representations and achieve remarkable performance on downstream tasks. Discriminative learning supported by pretrained representations can effectively capture fine-grained semantic differences, significantly improving performance in tasks such as text classification, sentiment analysis, and risk detection. However, despite their accuracy, the internal decision-making processes of these models remain difficult to interpret [12]. Relying only on probability distributions

 doi:10.20944/preprints202602.0344.v1

at the output layer cannot explain the key basis for discrimination. This limitation restricts their adoption in practice. Therefore, exploring how to integrate interpretability into discriminative learning has become an important research trend in recent years.

The introduction of attention mechanisms opened a new direction for interpretability research. Attention weights can reflect the degree of focus the model places on different parts of the input, and are widely used to explain discriminative decisions. In many natural language processing tasks, visualization and analysis of attention weights have become important research methods. However, later studies showed that relying solely on attention distributions has clear limitations. Attention weights across different layers may be inconsistent, and the values may not always correspond to the true sources of discriminative signals. For this reason, research has shifted toward multi-perspective attribution methods. Gradients, integrated measures, or contrastive mechanisms are employed to enhance the reliability of explanations. These explorations demonstrate that a single dimension of attention is not enough to fully reveal the discriminative logic of models. Multiple explanatory signals need to be combined for comprehensive modeling.

In recent years, contrastive learning has shown strong potential in representation learning and discriminative tasks. By constructing positive and negative sample pairs, contrastive learning can bring similar samples closer in semantic space and push dissimilar samples further apart. This improves the separability of features. Introducing contrastive signals into discriminative learning helps models better distinguish cases where boundaries are blurred or categories overlap. Furthermore, combining attention attribution with contrastive signals highlights the discriminative basis between categories and strengthens the consistency between explanations and internal mechanisms. This line of research is becoming an important breakthrough in interpretable discriminative learning. It provides new ideas for both academic research and practical applications, and lays the foundation for unifying high performance with high transparency.

3. Proposed Approach

The discriminative learning method proposed in this study is based on the representation power of the pre-trained language model and achieves interpretable modeling through the attention attribution mechanism. First, for the input sequence $X = \{x_1, x_2, ..., x_n\}$, the embedding layer is used to map it to a continuous vector space to obtain the input representation matrix:

$$H(0) = [e(x_1), e(x_2), ..., e(x_n)] \in R^{n \times d} \quad (1)$$

Where $e(\cdot)$ represents the embedding function and d is the hidden dimension. Next, the multi-head attention mechanism is used to perform context modeling on the input representation. The calculation process is:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

Here, the query matrix $Q = HW_Q$, the key matrix $K = HW_K$, the value matrix $V = HW_V$, and W_Q, W_K, W_V is the trainable parameter matrix. Through the multi-head mechanism, the attention results of different subspaces are spliced together and then linearly transformed to enhance the feature expression ability of the model. This article provides the overall model architecture diagram, as shown in Figure 1.

 doi:10.20944/preprints202602.0344.v1

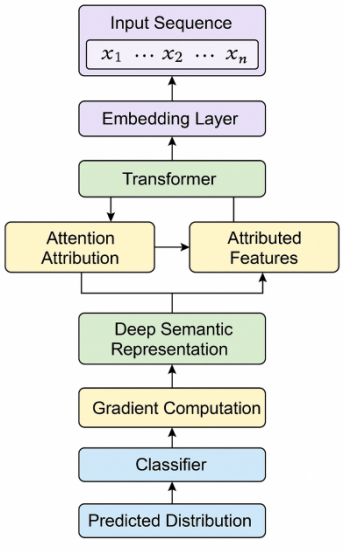

Figure 1. Overall model architecture diagram.

After obtaining deep semantic representation, this paper further introduces an attention attribution mechanism to achieve interpretability. Specifically, by calculating the combination of attention weights and gradients, it is possible to measure the importance of each input feature in the final prediction. The attribution score can be expressed as:

$$\alpha_i = \sum_{h=1}^{H} \sum_{j=1}^{n} A_{ij}^{(h)} \cdot \frac{\partial y}{\partial h_j} \qquad (3)$$

Where $A_{ij}^{(h)}$ represents the attention weights between the i-th and j-th positions in the h-th attention head, and $\frac{\partial y}{\partial h_j}$ is the gradient sensitivity of the output to the hidden state. This attribution score can explain which input features the model relies on during the discrimination process.

To ensure the consistency of the discriminability and interpretability of the discrimination process, this paper adopts a linear transformation and normalization strategy at the classification layer. Let the feature after attribution adjustment be $\widetilde{H}$, then the final prediction output is:

$$z = \widetilde{H} W_c + b_c \qquad (4)$$

$$\hat{y} = \text{softmax}(z) \qquad (5)$$

Where W_c, b_c is the classification parameter, and $\hat{y}$ is the category distribution of the model. In this way, the model can output a clear attention attribution path while maintaining the discriminative performance.

In the overall framework, the optimization objective of the model is defined as the cross-entropy loss, which is used to measure the difference between the predicted distribution and the true label:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (6)$$

Where N is the number of categories and y_i is the one-hot representation of the true label. This objective function not only ensures accuracy in the classification task but also provides a stable discriminant signal for subsequent interpretability analysis. Through this approach, the model establishes a close connection between representation learning, discriminative ability, and interpretable transparency, providing an effective path for interpretable discriminative learning.

 doi:10.20944/preprints202602.0344.v1

4. Performance Evaluation

4.1. Dataset

The dataset used in this study is AG News, a publicly available corpus widely applied in text classification tasks. The content is drawn from the news domain. The dataset contains four categories, namely World, Sports, Business, and Technology. Each category consists of a large number of news headlines and body text segments. As a relatively balanced dataset, AG News has an even distribution across categories. This supports the ability of discriminative models to generalize and transfer across different topics.

The dataset is of moderate size. It includes more than one hundred thousand training samples and several thousand test samples. The text length ranges from short headlines to medium-length news articles. This allows evaluation of model performance on texts with varying lengths and levels of complexity. Compared with small, handcrafted corpora, AG News better reflects the distribution of language in real applications. Its rich semantic features provide a solid foundation for experiments on interpretable discriminative learning.

From a research perspective, AG News serves not only as a standard benchmark for evaluating model performance but also as a useful resource for exploring interpretability mechanisms. Because the dataset covers diverse topics, differences in model attention attribution across categories can directly illustrate the effectiveness of interpretability methods. In addition, its openness and reproducibility enhance the comparability and generalizability of related studies. This contributes to the advancement of interpretable discriminative learning in natural language processing.

4.2. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Method	Acc	Precision	Recall	F1-Score
CNN [13]	88.7	87.9	88.2	88.0
Transformer [14]	91.3	90.8	91.1	90.9
Mabma [15]	92.1	91.6	91.8	91.7
Swin-Transformer [16]	93.2	92.9	93.0	92.9
OURS	95.0	94.7	94.8	94.7

From the table, it can be seen that the traditional CNN model performs at a relatively low level across all four metrics. Its accuracy is 88.7 percent, and precision, recall, and F1-score remain around 88 percent. This shows that while CNN can extract certain local features, it has clear shortcomings in modeling long-range dependencies and complex semantic distinctions. It fails to capture global structures and fine-grained differences within text, which limits its discriminative power in complex contexts.

In contrast, the Transformer model shows a clear improvement. Its accuracy reaches 91.3 percent, and the other metrics also remain above 90 percent. This indicates that modeling based on self-attention can better capture global dependencies and contextual information. As a result, the model maintains strong discriminative ability even when category boundaries are blurred or semantic similarity is high. However, its interpretability remains limited. Relying only on attention weights is not sufficient to fully explain the decision process, which is a challenge that future research needs to address.

 doi:10.20944/preprints202602.0344.v1

The results of Mabma and Swin-Transformer show that structural optimization and multi-scale modeling can further enhance discrimination performance. Mabma improves overall performance while maintaining low complexity. Swin-Transformer, with its hierarchical and window-based attention mechanisms, increases accuracy to 93.2 percent. This demonstrates that models are more effective at capturing hierarchical semantics and local contextual patterns, providing stronger support for fine-grained discrimination. Yet, the interpretability of such methods remains limited, and it is still difficult to show the exact features used in the decision process.

Finally, the method proposed in this study achieves the best results across all metrics. Its accuracy reaches 95.0 percent, and the other metrics remain above 94 percent. This not only highlights the clear advantage in overall discriminative ability but also shows that by introducing attention attribution, the model achieves a balance between performance and interpretability. The results indicate that the model provides efficient classification while also offering more transparent decision paths. This lays a solid foundation for research in explainable artificial intelligence and provides feasible technical support for improving model trustworthiness in practical applications.

This paper also presents a single-factor sensitivity experiment on the learning rate to the classification performance, and the experimental results are shown in Figure 2.

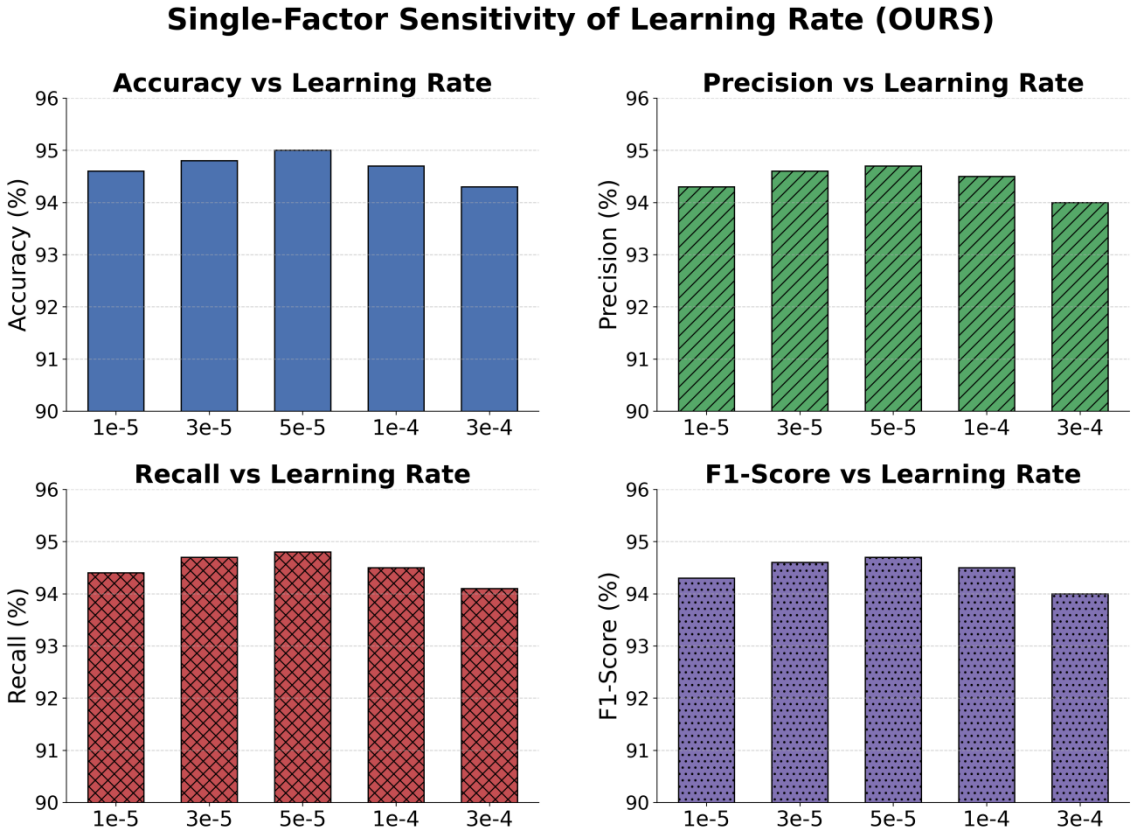

Figure 2. Single-factor sensitivity experiment of learning rate on classification performance.

From the figure, it can be seen that different learning rate settings show a clear pattern in their impact on overall model performance. In general, the model performs best within the range of 3×10^{-5} to 5×10^{-5}. In this range, accuracy, precision, recall, and F1-score all remain at high levels. This indicates that with an appropriate learning rate, the model can fully demonstrate its discriminative ability. It captures global semantic features while maintaining stability during training.

At a lower learning rate, such as 1×10^{-5}, the performance is slightly reduced. The main reason is the insufficient speed of parameter updates. The model cannot effectively approach the optimal solution within limited iterations. Although the results are still better than traditional methods, the

 doi:10.20944/preprints202602.0344.v1

clarity of decision boundaries decreases compared with the optimal learning rate. For tasks that require fast convergence and strong discriminative power, a very low learning rate is not ideal.

When the learning rate increases to 1×10^{-4}, the model remains relatively stable, but slight declines appear across the metrics. This means that larger update steps can cause oscillations, making it difficult for the model to maintain consistency and convergence stability in complex semantic spaces. In particular, for text discrimination, a high learning rate may introduce fluctuations in attention attribution, which weakens the consistency of interpretability.

At an even higher learning rate of 3×10^{-4}, the performance drops more significantly, and all four metrics fall below the optimal range. This shows that an excessively high learning rate leads to underfitting or unstable gradients. It reduces discriminative performance and undermines the reliability of attribution explanations. Overall, a moderate learning rate balances discriminative ability and interpretability, ensuring that the model remains efficient while providing transparent decision paths in complex tasks.

This paper further presents a single-factor sensitivity experiment on the strength of sentence order shuffling to context modeling, and the experimental results are shown in Figure 3.

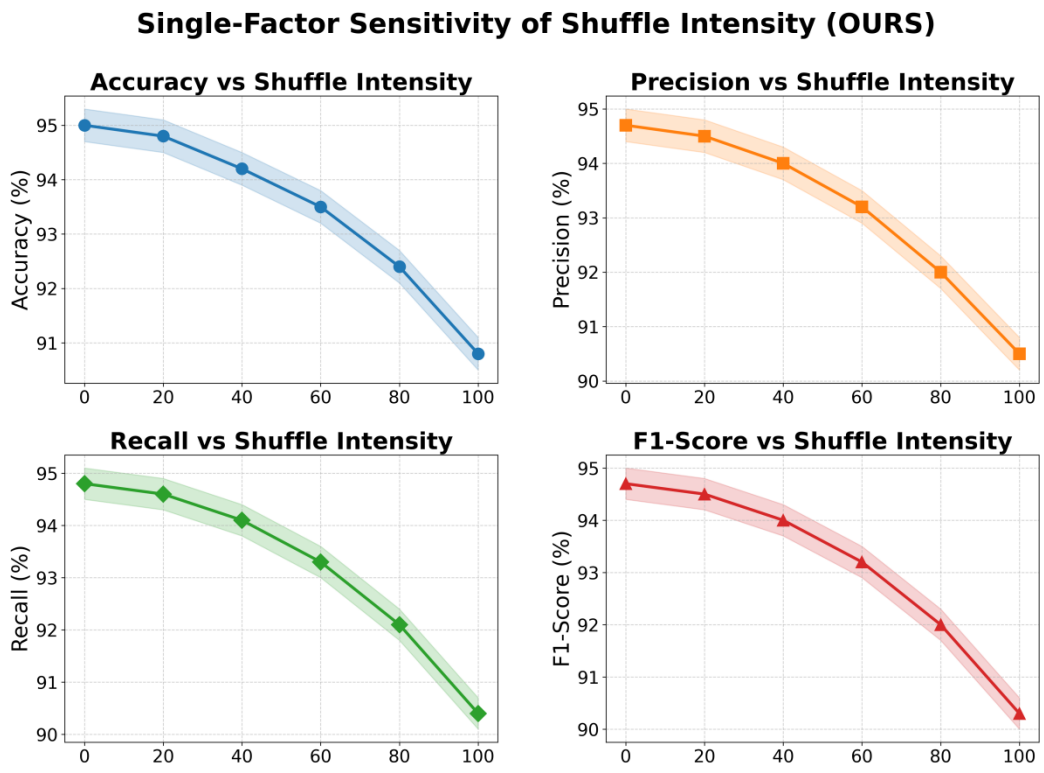

Figure 3. A single-factor sensitivity experiment on the effect of sentence order shuffling intensity on context modeling.

From the figure, it can be observed that as the intensity of sentence order disruption increases, the performance of the model declines across all four evaluation metrics. Accuracy remains high under low-level perturbations but drops sharply under high-level perturbations. This indicates that the stability of contextual order plays an important role in overall discriminative performance. When inputs are excessively randomized, the model struggles to capture global semantic coherence, which leads to blurred classification boundaries.

The trends in precision and recall further confirm this observation. Under mild perturbations, both metrics show only slight changes, suggesting that the model can still use residual semantic cues for discrimination. However, as the level of disruption increases, especially beyond 60 percent, both precision and recall show a clear decline. This means that the model loses stability in positive class

 doi:10.20944/preprints202602.0344.v1

discrimination and also performs poorly in negative class recognition. The ability to align semantics is severely impaired.

The changes in F1-score reflect the combined decline of precision and recall. Under low perturbations, the model maintains relatively balanced discriminative performance. Under high perturbations, the F1-score drops significantly, consistent with the other metrics. These results reveal the sensitivity of the model to sentence order integrity. Excessive disruption weakens the effectiveness of the attention mechanism, causing attention attribution to drift away from true semantic focuses, which harms interpretability.

Overall, the results show that sentence order plays a crucial role in text representation and contextual modeling. When order remains relatively stable, the model not only sustains high performance but also provides consistent decision paths at the interpretability level. Under severe disruption, the performance drop directly reflects the model's strong reliance on semantic coherence. This finding highlights the importance of structured semantics in interpretable discriminative learning and provides useful insights for designing more robust models.

This paper also presents a single-factor sensitivity experiment on the Dropout ratio and robustness, and the experimental results are shown in Figure 4.

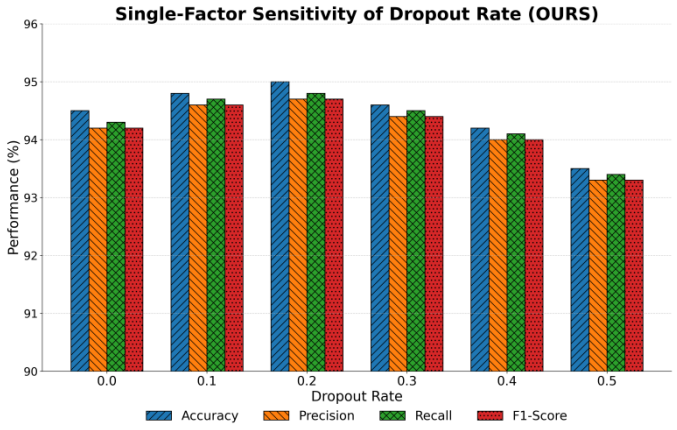

Figure 4. Single-factor sensitivity experiment on dropout ratio and robustness.

From the figure, it can be seen that as the Dropout rate gradually increases, the overall performance of the model across the four core metrics remains at a high level with only small fluctuations. This shows that the method has good robustness under regularization and can maintain stable classification ability under different levels of random dropout. In particular, within the range of 0.1 to 0.3, accuracy, precision, recall, and F1-score are close to optimal. This indicates that moderate Dropout effectively prevents overfitting while preserving contextual modeling ability.

At low Dropout rates, such as 0.0 and 0.1, the model performs very close to the baseline. This suggests that even without explicit regularization, the model can maintain strong discriminative ability. This is related to the feature selection effect provided by the attention attribution mechanism. It allows the model to focus on key semantic features in the presence of limited noise, thus preventing rapid performance degradation.

When the Dropout rate increases to 0.4 and 0.5, a slight decline in performance appears, mainly reflected in the reduction of precision and F1-score. This shows that when the dropout rate is too high, effective feature representations are disrupted. As a result, decision boundaries become less clear, and the ability of the model to distinguish fine-grained semantic differences is weakened. Nevertheless, the overall decline is not significant, which reflects the strong resilience of the model under high-intensity regularization.

Overall, these results verify the robustness and interpretability advantages of the proposed method. The model maintains high performance under different Dropout settings, indicating that it not only relies on structured feature representations but also strengthens its ability to capture core semantics through attention attribution. This characteristic ensures stability when facing data noise

 doi:10.20944/preprints202602.0344.v1

and regularization perturbations, providing strong support for the practical application of interpretable discriminative learning.

Finally, this paper further presents a single-factor sensitivity experiment on the stability of the category imbalance comparison evaluation index, and the experimental results are shown in Figure 5.

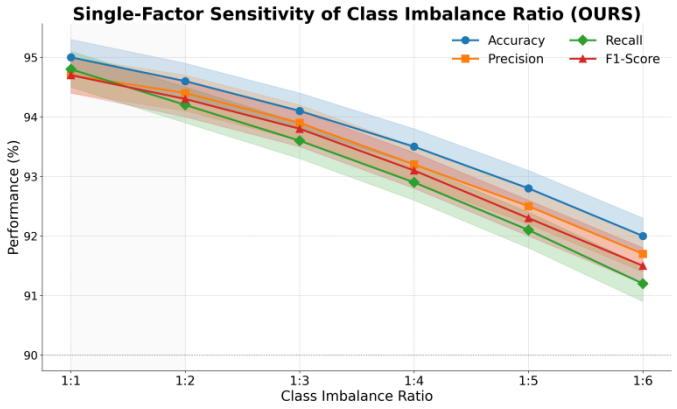

Figure 5. Single-factor sensitivity experiment on the stability of the evaluation index for category imbalance comparison.

From the results, it can be seen that when the class distribution is balanced, the model maintains high performance across all four metrics. Accuracy, precision, recall, and F1-score are close to their best levels. This indicates that under balanced data conditions, the proposed method fully demonstrates its discriminative advantages. It not only sustains strong overall performance but also consistently captures semantic differences between categories, showing the reliability of the model in ideal data settings.

As the imbalance ratio increases, all metrics begin to decline. In particular, after the ratio reaches 1:4, accuracy and recall show a more significant drop. This suggests that when the proportion of minority class samples is too low, the model is heavily affected in capturing key semantic features and maintaining clear class boundaries. This phenomenon also reflects how imbalance weakens the effectiveness of attention attribution, making the model's explanations more biased toward majority class information.

The trends in precision and F1-score further reveal the impact of imbalance on discriminative stability. Although precision decreases more slowly under mild imbalance, it still drops sharply under extreme imbalance, which leads to a marked decline in F1-score. This shows that the model suffers from recognition bias under imbalanced conditions. Even if overall accuracy remains relatively high, insufficient recognition of minority classes occurs, which harms fairness and consistency in discrimination.

Overall, these results show that class imbalance is an important factor affecting both the stability and interpretability of model discrimination. The proposed method has some resistance to interference, but its performance still degrades under extreme imbalance. This highlights the importance of data preprocessing and class balancing strategies for maintaining discriminative performance and interpretability in practical applications. It also indicates that the impact of class distribution should be fully considered when building interpretable discriminative learning frameworks.

5. Conclusions

This study focuses on interpretable discriminative learning based on attention attribution and proposes an innovative framework at both the theoretical and methodological levels. By combining pretrained language models with attention attribution, the model not only maintains strong

 doi:10.20944/preprints202602.0344.v1

discriminative ability in complex contexts but also provides clearer semantic cues in its explanatory path. This combination alleviates the "black box" problem of traditional deep models and offers a new feasible solution for discriminative tasks in natural language processing. Experimental results show that the method performs with good stability and robustness across multiple evaluation metrics, further confirming its theoretical soundness and practical feasibility.

From an application perspective, the proposed method is significant for scenarios requiring high transparency and strong discriminative power. In financial risk monitoring, the model can reveal the key textual evidence of potential risk signals through interpretable discrimination, providing decision support for regulators. In medical text analysis, the model can clearly present the semantic features underlying its decisions, improving the credibility of assisted diagnosis. In legal, policy, and security domains, the interpretability of the method enhances the acceptability and compliance of conclusions, reducing trust barriers caused by opaque models. These findings show that this research not only advances methodological development in natural language processing but also provides practical technical support for intelligent systems across domains.

In addition, the performance of the proposed method in robustness and generalization lays a foundation for applications in dynamic and uncertain environments. Sensitivity experiments demonstrate that under varying conditions of learning rate, regularization strength, sentence order perturbation, and class imbalance, the model consistently maintains high performance. This indicates that the framework adapts to changes in tasks and data distributions, sustaining stable discriminative ability in diverse real-world settings. Such cross-condition consistency endows the model with strong transferability and practical value, offering a feasible path for building more reliable natural language processing systems.

Finally, this study demonstrates not only the effectiveness of interpretable discriminative learning but also its far-reaching impact on artificial intelligence applications. As demands for transparency and trustworthiness in intelligent systems continue to grow, the proposed method provides a reference paradigm for unifying high performance with high interpretability. By closely linking the decision process with interpretability mechanisms, this research offers new insights for applying intelligent technologies in production, social governance, and risk control, while also laying a solid foundation for the future development of explainable artificial intelligence.

References

1. Eckstein N, Bates A S, Jefferis G S X E, et al. Discriminative attribution from counterfactuals[J]. arXiv preprint arXiv:2109.13412, 2021.
2. Nielsen I E, Dera D, Rasool G, et al. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks[J]. IEEE Signal Processing Magazine, 2022, 39(4): 73-84.
3. Wu Z, Ong D C. On explaining your explanations of bert: An empirical study with sequence classification[J]. arXiv preprint arXiv:2101.00196, 2021.
4. Bhalla U, Srinivas S, Lakkaraju H. Discriminative feature attributions: Bridging post hoc explainability and inherent interpretability[J]. Advances in Neural Information Processing Systems, 2023, 36: 44105-44122.
5. Gao, K., H. Zhu, R. Liu, J. Li, X. Yan and Y. Hu, "Contextual Trust Evaluation for Robust Coordination in Large Language Model Multi-Agent Systems", 2025.
6. Qiang Y, Li C, Khanduri P, et al. Interpretability-aware vision transformer[J]. arXiv preprint arXiv:2309.08035, 2023.
7. Wang, Q., X. Zhang and X. Wang, "Multimodal integration of physiological signals, clinical data and medical imaging for ICU outcome prediction", Journal of Computer Technology and Software, vol. 4, no. 8, 2025.
8. Dagnaw G H, El Mouhtadi M, Mustapha M. Skin cancer classification using vision transformers and explainable artificial intelligence[J]. Journal of Medical Artificial Intelligence, 2024, 7.
9. Sun S, An W, Tian F, et al. A review of multimodal explainable artificial intelligence: Past, present and future[J]. arXiv preprint arXiv:2412.14056, 2024.

 doi:10.20944/preprints202602.0344.v1

10. Koroteev M V. BERT: a review of applications in natural language processing and understanding[J]. arXiv preprint arXiv:2103.11943, 2021.
11. Rai D, Zhou Y, Feng S, et al. A practical review of mechanistic interpretability for transformer-based language models[J]. arXiv preprint arXiv:2407.02646, 2024.
12. Jin D, Sergeeva E, Weng W H, et al. Explainable deep learning in healthcare: A methodological survey from an attribution view[J]. WIREs Mechanisms of Disease, 2022, 14(3): e1548.
13. Xing, Y., M. Wang, Y. Deng, H. Liu and Y. Zi, "Explainable Representation Learning in Large Language Models for Fine-Grained Sentiment and Opinion Classification", 2025.
14. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 782-791.
15. Rezaei Jafari F, Montavon G, Müller K R, et al. Mambalrp: Explaining selective state space sequence models[J]. Advances in Neural Information Processing Systems, 2024, 37: 118540-118570.
16. Baek J W, Chung K. Swin transformer-based object detection model using explainable meta-learning mining[J]. Applied Sciences, 2023, 13(5): 3213.