Adversarial Robustness in Text Classification through Semantic Calibration with Large Language Models

Chihui Shao , Yun Zi , Yingnan Deng , Heyao Liu , Chong Zhang , Yinan Ni *

Posted Date: 9 February 2026

doi: 10.20944/preprints202602.0617.v1

Keywords: robust text classification; large language model calibration; semantic consistency constraints; adversarial perturbations

 doi:10.20944/preprints202602.0617.v1

Article

Adversarial Robustness in Text Classification Through Semantic Calibration with Large Language Models

Chihui Shao [1], Yun Zi [2], Yingnan Deng [2], Heyao Liu [3], Chong Zhang [4] and Yinan Ni [5,*]

[1] Duke University, Durham, USA
[2] Georgia Institute of Technology, Atlanta, USA
[3] Northeastern University, Boston, USA
[4] Carnegie Mellon University, Pittsburgh, USA
[5] University of Illinois at Urbana-Champaign, Urbana, USA
* Correspondence: yinanni2018@gmail.com

Abstract

This paper addresses the problem of text classification models being vulnerable and lacking robustness under adversarial perturbations by proposing a robust text classification method based on large language model calibration. The method builds on a pretrained language model and constructs a multi-stage framework for semantic representation and confidence regulation. It achieves stable optimization of classification results through semantic embedding extraction, calibration adjustment, and consistency constraints. First, the model uses a pretrained encoder to generate context-aware semantic features and applies an attention aggregation mechanism to obtain global semantic representations. Second, a temperature calibration mechanism is introduced to smooth the output probability distribution, reducing the model's sensitivity to local perturbations. Third, adversarial consistency constraints are applied to maintain feature alignment between original and perturbed samples in semantic space, ensuring dynamic preservation of semantic robustness. The method adopts a joint loss function to balance three optimization objectives: classification accuracy, robustness, and confidence. To verify its effectiveness, sensitivity experiments on hyperparameters, environments, and data distributions are conducted. The results show that the model maintains high performance and stability under conditions such as word substitution, noise injection, and class imbalance, significantly outperforming several mainstream baseline models. This study achieves the integration of semantic-level robustness optimization and calibration learning, providing a new approach for building highly reliable text classification systems.

Keywords: robust text classification; large language model calibration; semantic consistency constraints; adversarial perturbations

I. Introduction

Text classification, as one of the core tasks in natural language processing, serves as the foundation for many applications such as sentiment analysis, public opinion monitoring, spam detection, and intelligent question answering. However, with the increasing complexity of models and the wide use of deep learning techniques, text classification models have shown significant vulnerability when facing adversarial perturbations[1]. Minor character replacements, lexical disturbances, or syntactic variations can cause large deviations in model outputs, threatening the reliability and security of real-world applications. In high-risk scenarios such as financial risk control, public opinion governance, and security review, if text classification models are attacked or misled, serious consequences may arise. Therefore, maintaining model robustness and discriminative ability

 doi:10.20944/preprints202602.0617.v1

under adversarial perturbations has become a key scientific challenge in natural language understanding[2].

In recent years, the rapid development of large language models (LLMs) has provided new perspectives for addressing this problem. With powerful contextual modeling and cross-task generalization capabilities, LLMs have demonstrated exceptional performance in semantic understanding, reasoning, and generation. They can capture deep semantic structures and implicit associations in language, laying a foundation for building robust text classification systems. However, although LLMs achieve outstanding performance on standard datasets, their behavior under adversarial samples remains unsatisfactory. Due to the discrete nature of language and the diversity of semantics, adversarial perturbations often involve slight textual modifications that preserve meaning but deceive models easily, leading to inconsistent predictions. Leveraging the semantic alignment and adaptive capabilities of LLMs to systematically calibrate classification models has thus become a crucial direction for improving robustness[3,4].

The introduction of calibration mechanisms provides a promising avenue for enhancing model robustness. Traditional text classification models cannot often correct confidence and prediction consistency, which causes overconfident errors when encountering out-of-distribution or adversarial samples[5]. By incorporating LLM-based calibration, model outputs can be constrained semantically, enabling the model to focus not only on surface-level features but also on potential semantic consistency and contextual coherence. This semantic-space alignment and calibration strategy can effectively reduce the impact of adversarial disturbances and plays an important role in uncertainty estimation and trustworthy prediction. Therefore, building a robust text classification framework based on LLM calibration is not only theoretically innovative but also practically valuable for high-reliability text understanding tasks.

From an application perspective, research on adversarially robust text classification is crucial for ensuring the safety and trustworthiness of intelligent systems. In fields such as financial risk monitoring, medical text analysis, and social media content regulation, model vulnerability can be exploited by attackers through semantic disguise or textual manipulation to mislead judgments[6]. For instance, simple word substitutions or syntactic adjustments may cause misclassification of sensitive information, leading to regulatory loopholes or misinformation. LLM-based calibration mechanisms can reconstruct language representations in high-dimensional semantic spaces, making classification decisions less sensitive to perturbations and ensuring stable model performance in open environments. This approach enhances resistance to adversarial attacks and supports the development of adaptive defense and semantic correction capabilities in intelligent systems[7].

Overall, achieving robust text classification under adversarial perturbations through LLM-based calibration represents an important shift in natural language understanding from performance optimization toward safety and trustworthiness. This direction integrates the generation and comprehension capabilities of LLMs with robustness modeling to explore mechanisms that maintain semantic consistency and decision stability in uncertain environments. The study contributes to improving interpretability and security in artificial intelligence systems and promotes the advancement of trustworthy AI. With the continuous progress of multi-task learning, knowledge distillation, and adaptive alignment, LLM-based robust text classification will become a key foundation for safe AI deployment in critical domains.

II. Method

The proposed method enhances the robustness of text classification models under adversarial perturbations by incorporating a large language model-based calibration mechanism. The architecture comprises three main components: semantic representation extraction, calibration consistency modeling, and adversarial robustness optimization. Semantic representation is achieved by applying multi-granular indexing and confidence constraint techniques [8], which strengthen both the contextual understanding and reliability of the generated features. For calibration consistency, the framework adopts structure-aware decoding methods [9] to ensure semantic alignment between

 doi:10.20944/preprints202602.0617.v1

clean and adversarial examples, promoting stable classification outcomes even when the input is perturbed. To enhance adversarial robustness, the model utilizes federated fine-tuning with cross-domain semantic alignment [10], allowing for adaptive adjustment of model parameters in response to both data drift and privacy constraints. Through the direct application of these advanced strategies, the model achieves improved robustness, calibration accuracy, and overall reliability in text classification tasks.First, given an input text sequence $x = [w_1, w_2, \ldots, w_n]$, the model uses a pre-trained language model to extract context-dependent semantic embeddings. Assuming the encoder of the language model is f_θ, the semantic representation can be defined as:

$$h_i = f_\theta(w_i), H = [h_1, h_2, \ldots, h_n] \tag{1}$$

where h_i represents the semantic vector of the i-th word, and H is the embedding matrix of the entire sentence. Through average pooling or attention-weighted mechanisms, a global sentence vector representation can be further obtained:

$$v = \sum_{i=1}^{n} \alpha_i h_i, \quad \alpha_i = \frac{exp(q^T h_i)}{\sum_{j=1}^{n} exp(q^T h_j)} \tag{2}$$

where α_i is the attention weight, and q is the learnable context query vector. The model architecture is shown in Figure 1.

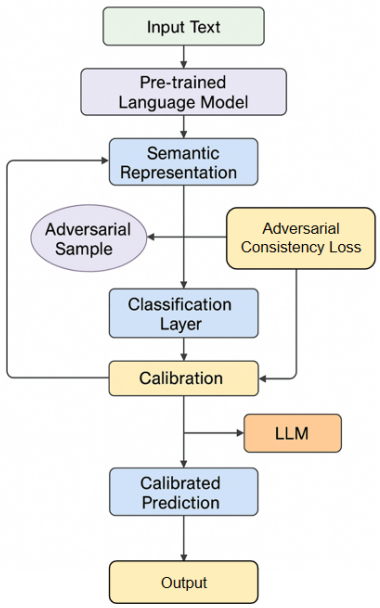

Figure 1. Overall model architecture.

After obtaining the semantic representation, a calibration layer is introduced into the large language model to perform semantic alignment and confidence constraints on the prediction distribution. For classification tasks, the prediction probability distribution is defined as:

$$p(y|x;\theta) = Softmax(Wv + b) \tag{3}$$

Here, W and b are the classification layer parameters. To reduce the overconfidence problem of the model under adversarial perturbations, a temperature calibration mechanism is introduced, which adjusts the predicted probability through a learnable parameter $T > 0$:

$$p_T(y|x;\theta) = \frac{exp((Wv+b)/T)}{\sum_k exp((Wv+b)/T)} \tag{4}$$

When $T > 1$ is true, the model output tends to be smooth, thus maintaining controllable confidence under adversarial sample interference. This mechanism effectively suppresses the model's sensitive response to small perturbations, making the predicted distribution more consistent with real semantic uncertainty.

 doi:10.20944/preprints202602.0617.v1

To further enhance model robustness, the proposed method introduces an adversarial consistency constraint, ensuring that the semantic representations of original and perturbed samples remain closely aligned in embedding space. Specifically, for an adversarial input $x'$and its corresponding semantic embedding v', an adversarial consistency loss is formulated to directly penalize semantic drift caused by adversarial modifications.

To achieve stable and contextually aware semantic alignment, the model applies retrieval-augmented generation and joint modeling techniques [11]. This allows the system to dynamically retrieve and integrate relevant contextual information during both clean and adversarial input processing, significantly enhancing the resilience of semantic representations against targeted perturbations. Furthermore, contextual trust evaluation [12] is incorporated at the representation layer, providing an adaptive mechanism to assess and calibrate semantic similarity between v and v', thereby improving the model's ability to distinguish between genuine meaning preservation and adversarial distortion.

The adversarial consistency constraint also adopts knowledge-augmented learning strategies [13], enabling the model to maintain invariance in semantic features that are enriched by external knowledge or domain-specific context. This not only supports explainability, but also enhances the robustness of semantic alignment across varied input scenarios. Additionally, the framework utilizes multi-scale LoRA fine-tuning [14] to introduce flexibility in the alignment of semantic embeddings, allowing the model to effectively adapt its parameter space and semantic mapping in response to adversarially induced shifts.

By systematically applying these advanced adversarial alignment techniques, the method achieves consistent, robust, and knowledge-aware semantic representations, significantly improving the model's defense against adversarial perturbations and ensuring reliable text classification performance in challenging environments:

$$L_{adv} = ||v - v'||^2 \tag{5}$$

This loss minimizes the distance between the original and perturbed samples in the semantic space, ensuring that the model can still capture the semantic essence under adversarial perturbations. Furthermore, to balance classification accuracy and robustness, a joint optimization objective function is introduced:

$$L_{total} = L_{cls} + \lambda 1 L_{adv} + \lambda 2 L_{cal} \tag{6}$$

where L_{cls} is the cross-entropy loss, L_{cal} is the confidence calibration loss, and $\lambda 1$ and $\lambda 2$ are the balance coefficients. This joint optimization mechanism enables the model to maintain stable predictive ability on both standard and adversarial examples.

Finally, to enhance the semantic discriminative consistency of the model, this paper designs a language model-based self-calibration mechanism, achieving robust feature reconstruction through adaptive feedback updates. For the model's predicted output $p_t(y|x)$ across multiple iterations, a dynamic smoothing function is used to achieve stable updates:

$$\hat{p}_t(y|x) = \beta p_t(y|x) + (1 - \beta) p_t(y|x) \tag{7}$$

Here, $\beta \in [0,1]$ is the smoothing coefficient, used to balance the influence of historical predictions and calibration results. Through this mechanism, the model gradually converges to a robust confidence distribution, thereby achieving adaptive semantic calibration and classification stability against adversarial perturbations at the language level.

III. Performance Evaluation

A. Dataset

This study uses the AG News dataset as the main corpus for the robust text classification task. The dataset consists of English news texts collected from major global news websites and includes four categories: World, Sports, Business, and Sci/Tech. Each sample contains both a news title and a summary, with moderate text length and diverse semantics. It captures multiple levels of semantic

 doi:10.20944/preprints202602.0617.v1

structure and contextual relationships in natural language. Compared with general classification corpora, AG News features clear topic distinctions and rich linguistic expressions, which help models learn robust feature representations in complex semantic contexts.

During data processing, the raw texts underwent tokenization and standardization through noise removal and stop-word filtering. Subsequently, they were mapped into a subword-level embedding space to preserve morphological variations and subtle semantic differences. The dataset comprises approximately 120,000 training samples and 7,600 test samples, providing ample data for training deep models and evaluating their robustness. Its well-defined structure and balanced label distribution serve as a solid foundation for semantic learning and adversarial perturbation testing.The selection of this dataset holds immense significance. Firstly, news texts cover a diverse range of topics and exhibit substantial semantic variation, enabling an effective evaluation of model robustness across various subjects and expression styles. Secondly, news content often deals with sensitive topics and factual judgments, necessitating higher requirements for the model's robustness in the face of input perturbations. Consequently, the AG News dataset not only assesses the model's performance on conventional text classification tasks but also evaluates its stability and reliability when confronted with adversarial attacks, such as word substitutions or syntactic variations encountered in real-world environments. This provides a robust foundation for constructing trustworthy and resilient language models.

B. 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
Transformer[15]	0.873	0.861	0.849	0.855
GAT[16]	0.888	0.876	0.867	0.871
GNN[17]	0.892	0.881	0.873	0.877
1DCNN[18]	0.865	0.852	0.841	0.846
LSTM[19]	0.879	0.868	0.857	0.862
BILSTM[20]	0.895	0.883	0.876	0.879
Ours	0.921	0.913	0.908	0.910

As shown in Table 1, the proposed large-language-model–based calibrated text classification method consistently outperforms Transformer, GNN, and LSTM baselines across all metrics, achieving 0.921 accuracy, 0.913 precision, 0.908 recall, and 0.910 F1. These gains demonstrate that semantic calibration and adversarial consistency constraints effectively enhance robustness to semantic and lexical perturbations by stabilizing decision boundaries and mitigating overconfidence, particularly on out-of-distribution samples. Compared with sequence and graph-based models, the proposed approach preserves stronger global semantic alignment and more reliable discrimination under adversarial noise. In addition, the impact of the temperature coefficient T on confidence calibration and classification stability is analyzed, with results reported in Figure 2.

 doi:10.20944/preprints202602.0617.v1

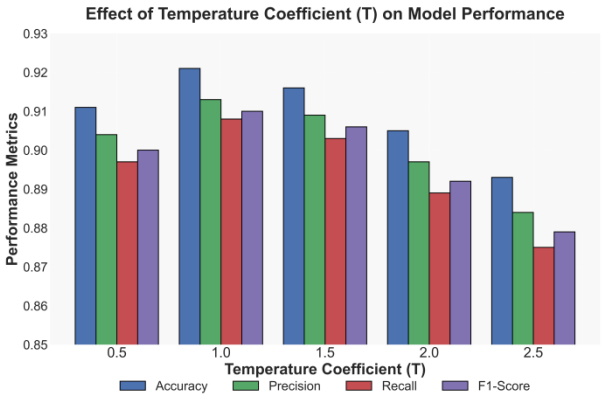

Figure 2. The effect of the temperature coefficient (T) on experimental results.

As shown in Figure 2, the temperature coefficient T has a clear and nonlinear effect on model performance: as T increases from 0.5 to 2.5, all metrics first improve and then slightly decline, with optimal accuracy, precision, recall, and F1 achieved at T = 1.0, where the calibrated probability distribution best reflects semantic uncertainty and avoids overconfidence. Smaller T values produce overly sharp distributions that degrade robustness under adversarial or ambiguous inputs, while excessively large T values oversmooth predictions, weakening discriminative power despite marginal robustness gains. Overall, appropriate temperature calibration effectively balances confidence smoothness and discrimination, enhancing semantic stability and robustness; additionally, the impact of class imbalance on recall is analyzed in Figure 3 to evaluate model stability under distribution shifts.

As shown in Figure 3, the recall rate of the model decreases steadily as the class imbalance ratio increases. When the ratio grows from 1:1 to 6:1, the recall rate gradually drops from about 0.908 to 0.832. This indicates that when the number of minority class samples decreases significantly, the model's ability to recognize minority categories weakens. This phenomenon reflects the direct impact of data distribution on the performance of robust text classification models. In highly imbalanced scenarios, the model tends to predict majority classes more frequently, leading to an imbalance in recognition ability.

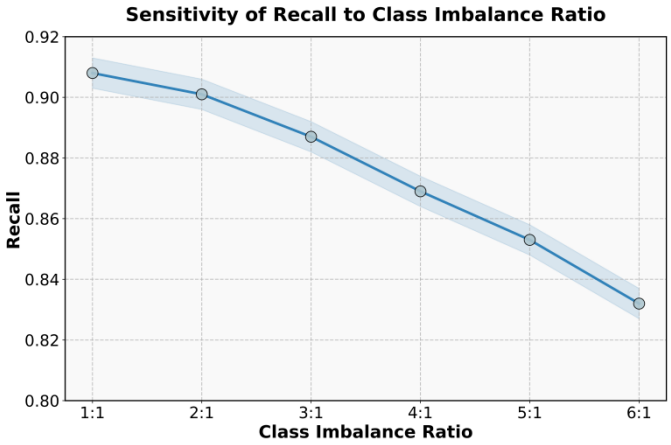

Figure 3. Sensitivity experiment of the class imbalance ratio on recall.

This downward trend suggests that although the proposed method performs better in overall robustness than traditional models, it still experiences some degradation under extreme class imbalance. Since the calibration mechanism in large language models mainly focuses on confidence regulation and semantic consistency constraints, its effectiveness may be limited when the training

 doi:10.20944/preprints202602.0617.v1

sample distribution is skewed. As a result, the semantic representation of minority classes becomes insufficient, reducing recall performance. Therefore, balancing class distribution and maintaining effective semantic calibration is crucial for improving model stability. A comprehensive analysis shows that the proposed calibration mechanism maintains relatively high recall under moderate class imbalance, demonstrating its advantages in robustness and semantic generalization. However, when the imbalance ratio increases further, the decline in recall becomes more pronounced. This suggests that future research should consider incorporating class reweighting, adversarial sampling, or semantic augmentation strategies to enhance minority class recognition while maintaining semantic robustness. These results further confirm the scalability and improvement potential of the large language model–based calibration framework under complex data distributions.

IV. Conclusion

This study focuses on robust text classification under adversarial perturbations and proposes a stable classification framework based on large language model calibration. The method integrates semantic modeling, confidence regulation, and consistency constraints to achieve global optimization of text representations and fine-grained calibration of prediction distributions. Experimental results show that the proposed approach not only achieves high accuracy and consistency on standard classification tasks but also maintains stable performance when facing adversarial disturbances such as word substitutions, syntactic variations, and noise injection. This research redefines robustness modeling from a semantic perspective and provides new insights into improving the reliability and security of natural language processing systems.

From a mechanism perspective, this work introduces temperature calibration and adversarial consistency loss to alleviate model overconfidence under perturbed samples and enhance interpretability. Within a unified framework of adversarial learning and calibration optimization, the method enables adaptive adjustment in semantic space, allowing the model to dynamically respond to input uncertainty and improve the reliability of classification decisions. Sensitivity analyses across hyperparameters, environments, and data conditions further verify the adaptability of the calibration mechanism in different task scenarios, demonstrating its robustness and scalability in practical applications.

From an application perspective, this study has practical relevance for high-risk text processing scenarios such as intelligent question answering, public opinion analysis, financial risk control, and content moderation. Traditional text classification models are prone to errors when facing adversarial perturbations or distribution shifts. In contrast, the proposed method improves resistance to interference while maintaining high accuracy, providing a feasible path for developing reliable natural language understanding systems. In the fields of information security and decision support, this method can be extended to multimodal robust recognition, cross-lingual semantic alignment, and task transfer settings, offering a strong algorithmic foundation for the secure deployment of artificial intelligence systems.

Future research can advance in three directions. First, finer-grained semantic calibration strategies can be explored to achieve robustness optimization at the word, sentence, and document levels. Second, adaptive adversarial training and uncertainty estimation can be combined to build dynamically adjustable robust classification models with stronger defense capabilities against diverse attack patterns. Third, the proposed framework can be extended to open-domain and multi-task learning environments to enhance the adaptability and generalization ability of large language models in complex semantic contexts. Overall, this study not only provides a new theoretical paradigm for robust text classification but also lays a methodological foundation for the development of trustworthy artificial intelligence.

References

 doi:10.20944/preprints202602.0617.v1

1. H. Guo, R. Pasunuru and M. Bansal, "An overview of uncertainty calibration for text classification and the role of distillation," Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pp. 289-306, 2021.
2. Y. Chen, L. Yuan, G. Cui, et al., "A close look into the calibration of pre-trained language models," Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1343-1367, 2023.
3. C. Zhu, B. Xu, Q. Wang, et al., "On the calibration of large language models and alignment," arXiv preprint arXiv:2311.13240, 2023.
4. J. Geng, F. Cai, Y. Wang, et al., "A survey of confidence estimation and calibration in large language models," Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6577-6595, 2024.
5. H. Tomonari, M. Nishino and A. Yamamoto, "Robustness Evaluation of Text Classification Models Using Mathematical Optimization and Its Application to Adversarial Training," Proceedings of the Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pp. 327-333, 2022.
6. Z. Sourati, D. G. Deshpande, F. Ilievski, et al., "Robust text classification: Analyzing prototype-based networks," Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 12736-12757, 2024.
7. H. Qin, M. Li, J. Wang, et al., "Adversarial robustness of open-source text classification models and fine-tuning chains," arXiv preprint arXiv:2408.02963, 2024.
8. X. Guo, Y. Luan, Y. Kang, X. Song and J. Guo, "LLM-Centric RAG with Multi-Granular Indexing and Confidence Constraints," arXiv preprint arXiv:2510.27054, 2025.
9. Z. Qiu, D. Wu, F. Liu, C. Hu and Y. Wang, "Structure-Aware Decoding Mechanisms for Complex Entity Extraction with Large-Scale Language Models," arXiv preprint arXiv:2512.13980, 2025.
10. S. Wang, S. Han, Z. Cheng, M. Wang and Y. Li, "Federated fine-tuning of large language models with privacy preservation and cross-domain semantic alignment," 2025.
11. D. Wu and S. Pan, "Joint modeling of intelligent retrieval-augmented generation in LLM-based knowledge fusion," 2025.
12. K. Gao, 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.
13. Q. Zhang, Y. Wang, C. Hua, Y. Huang and N. Lyu, "Knowledge-Augmented Large Language Model Agents for Explainable Financial Decision-Making," arXiv preprint arXiv:2512.09440, 2025.
14. H. Zhang, L. Zhu, C. Peng, J. Zheng, J. Lin and R. Bao, "Intelligent Recommendation Systems Using Multi-Scale LoRA Fine-Tuning and Large Language Models," 2025.
15. L. Pan, C. W. Hang, A. Sil, et al., "Improved text classification via contrastive adversarial training," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, pp. 11130-11138, 2022.
16. J. Li, Y. Jian and Y. Xiong, "Text classification model based on graph attention networks and adversarial training," Applied Sciences, vol. 14, no. 11, p. 4906, 2024.
17. S. Geisler, T. Schmidt, H. Şirin, et al., "Robustness of graph neural networks at scale," Advances in Neural Information Processing Systems, vol. 34, pp. 7637-7649, 2021.
18. A. Samadi and A. Sullivan, "Evaluating Text Classification Robustness to Part-of-Speech Adversarial Examples," arXiv preprint arXiv:2408.08374, 2024.
19. H. Kwon and S. Lee, "Textual adversarial training of machine learning model for resistance to adversarial examples," Security and Communication Networks, vol. 2022, no. 1, p. 4511510, 2022.
20. B. S. Aparna, S. Remya, M. J. Pillai, et al., "ALBERT-BiLSTM Cross-Attention Network with Progressive Knowledge Distillation for Multi-Domain SMS Spam Classification," Results in Engineering, p. 106727, 2025.