# Robust Sentiment Classification Under Noisy, Informal, and Adversarial Text Conditions

Xinjiao Sheng
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
xinjie.work@binghamton.edu

**Abstract**

Sentiment classification has become a central capability in large-scale computational social science, market intelligence, and content moderation systems. However, real-world deployment remains constrained by pervasive noisy input, nonstandard orthography, code-switching, informal registers, and deliberate adversarial manipulation. This paper develops a system-level analysis of robust sentiment classification under such conditions. Rather than treating robustness as a narrow model architecture problem, the discussion frames it as a property emerging from interactions among data governance, annotation infrastructure, training objectives, evaluation protocols, and deployment monitoring. The paper examines structural trade-offs between representational expressiveness and vulnerability to perturbation, between data augmentation and distributional drift, and between interpretability and performance under attack. It provides a conceptual taxonomy of robustness failures, including phonetic substitutions, Unicode confusables, syntactic paraphrases, and social-contextual ambiguity. The discussion further addresses fairness implications when robustness interventions systematically benefit majority dialects while marginalizing minority language varieties. A forward-looking perspective integrates dynamic attention mechanisms, supervised contrastive learning, adversarial training, and human-in-the-loop governance into a coherent system design. The analysis emphasizes that sustainable robustness cannot be achieved by model retraining alone; it requires institutional arrangements for continuous evaluation, shared threat intelligence, provenance tracking, and regulatory alignment. The paper concludes by outlining research directions that couple technical robustness with organizational accountability, thereby enabling sentiment classification systems to remain reliable, equitable, and auditable in contested information environments.

**Keywords**

sentiment classification, adversarial robustness, noisy text, informal language, data governance, fairness, socio-technical systems, model deployment.

## 1. Introduction

Sentiment classification has evolved from early lexicon-based pipelines to large pretrained language models capable of nuanced affective inference across domains and languages. Foundational surveys have mapped this evolution and emphasized the dependence of performance on the quality and representativeness of training data [1]. Subsequent synthesis expanded the scope to opinion targets, comparative expressions, and domain adaptation, establishing sentiment analysis as a mature subfield with substantial practical value [2]. The widespread adoption of transformer architectures further improved benchmark accuracy by capturing contextual dependencies at scale [3]. Yet these gains have not translated into consistent reliability in the noisy, informal, and adversarial text conditions that characterize

social media, messaging platforms, user-generated reviews, and multilingual online communities.

The gap between benchmark performance and real-world robustness arises from several structural sources. First, text in deployment environments frequently contains typographical errors, dialectal variation, emoji, nonstandard capitalization, and inconsistent punctuation. Second, informal registers often violate the syntactic and lexical assumptions embedded in curated evaluation corpora. Third, hostile actors deliberately craft inputs to induce misclassification, leveraging gradient information, paraphrasing tools, or character-level perturbations [4]. Adversarial attacks on discrete text have demonstrated that models can be forced to change their predictions by altering a single word or character [5]. Other attack families use language-model-guided substitution to produce semantically equivalent but syntactically different examples [6]. Systematic evaluation has shown that standard fine-tuned models remain vulnerable to simple synonym replacements and rule-based transformations, undermining confidence in their deployment for security-sensitive tasks [7].

This paper adopts a systems perspective rather than a narrow algorithmic lens. Robustness is treated as an emergent property of the entire socio-technical infrastructure, including data collection, annotation, pretraining, fine-tuning, evaluation, deployment monitoring, and governance. The central argument is that design choices in any one of these layers can either amplify or mitigate vulnerabilities introduced in another. A model trained on clean newswire text may achieve high validation accuracy but fail catastrophically when exposed to misspelled social media posts. Similarly, a data augmentation strategy that oversamples synthetic perturbations can improve robustness to those specific perturbations while introducing biases that harm accuracy on naturally occurring minority dialects. The paper therefore emphasizes structural trade-offs rather than isolated technical fixes.

The remainder of this paper is organized as follows. Section 2 defines the core problem space, distinguishing among noise, informality, and adversarial perturbation while showing their operational overlaps. Section 3 examines architectural foundations, including character-level, convolutional, and attention-based encoders, and discusses how representational choices shape vulnerability. Section 4 addresses data governance and annotation infrastructure, focusing on provenance, label quality, and the construction of robustness evaluation suites. Section 5 analyzes training strategies, including adversarial training, contrastive objectives, and dynamic attention mechanisms. Section 6 considers evaluation, fairness, and deployment implications, with particular attention to distributional shift and auditability. Section 7 discusses system-level trade-offs and sustainability. Section 8 outlines governance and policy directions, and Section 9 concludes.

## 2. The Problem Space: Noise, Informality, and Adversarial Perturbation

Noise, informality, and adversarial perturbation are often conflated in the robustness literature, but they have distinct generative mechanisms and require different governance responses. Noise is typically unintentional, arising from keyboard errors, automatic speech recognition mistakes, optical character recognition failures, or low-resource language transfer effects. Informal language is intentional but nonstandard, reflecting social identity, brevity norms, or creative orthography. Adversarial perturbation is intentional and hostile, designed to exploit model weaknesses. Despite these differences, all three conditions converge on a common challenge: the distribution of deployment text diverges systematically from the distribution of training and validation text.

Character-level convolutional networks were among the first architectures to address spelling variation and morphologically rich noisy text by learning subword representations without relying on a fixed vocabulary [10]. Sentence-level convolutional models further demonstrated that local n-gram features can capture sentiment-bearing expressions even when overall syntax is degraded [11]. These approaches provide a degree of robustness to surface-level noise because they do not require exact token identity. However, they are less effective against paraphrase-level attacks that preserve characters but alter syntactic or semantic structure.

The introduction of self-attention mechanisms changed the robustness landscape by enabling models to weigh contextual information globally across a sequence [12]. Attention-based encoders can, in principle, disambiguate misspelled words by attending to surrounding context. Yet the same flexibility also creates attack surfaces. Gradient-based methods can identify high-impact positions in the input and substitute them with carefully selected alternatives [5]. Language-model-guided attacks generate fluent adversarial examples that are difficult for humans to detect but highly effective at flipping predictions [6]. Benchmark evaluations have shown that standard transformer classifiers often drop below random performance when attacked with simple word-level substitutions, revealing a sharp mismatch between in-distribution accuracy and adversarial resilience [7].

Adversarial training offers one response by augmenting training data with perturbed examples. Early work demonstrated that adversarial regularization can improve stability in text classification under semi-supervised settings [9]. More recent adversarial training paradigms formulate robustness as a min-max optimization problem and show that models can be hardened against a range of perturbation budgets [18]. However, such hardening is not free. It often reduces clean accuracy, increases training cost, and can lead to overfitting to specific attack algorithms. Moreover, robustness against one family of attacks does not transfer automatically to another, a phenomenon well documented in both image and text domains [16].

A crucial distinction must be made between robustness to naturally occurring distributional shift and robustness to worst-case adversarial manipulation. Natural adversarial examples include ambiguous phrasing, negation, sarcasm, and domain-specific terminology that arise without malicious intent [24]. These examples are less dramatic than deliberate attacks but more prevalent in deployment. Systems must therefore balance defense against low-probability high-impact attacks with defense against high-probability low-impact noise. A governance framework that ignores either source will yield brittle systems in practice.

## 3. Architectural Foundations for Robust Sentiment Classification

The choice of text representation fundamentally constrains downstream robustness. Word-level tokenization assumes a fixed vocabulary and is highly sensitive to out-of-vocabulary items. Character-level models mitigate this by operating over Unicode code points, which enables them to process misspellings and obfuscations that would otherwise be invisible [10]. Hybrid models combine character and word information to achieve both semantic expressiveness and surface robustness. Nevertheless, character-level representations can be exploited through visually confusable characters or Unicode control characters that are visually indistinguishable from standard letters [19]. The attack surface thus migrates rather than disappears.

Attention-based architectures introduced by the transformer model offer powerful contextual encoding but require careful fine-tuning for classification tasks [12]. A number of fine-tuning procedures have been proposed to stabilize transformer-based text classification, including layer-wise learning rate decay and adversarial pretraining objectives [14]. Pre-trained encoders such as ELECTRA further improve efficiency and robustness by learning to distinguish real tokens from generated replacements, a task that naturally increases sensitivity to token-level corruption [13]. However, these advances do not automatically translate into robustness against adversarial paraphrase attacks. The expressiveness of attention mechanisms can be diverted by carefully designed distractors or by exploiting spurious correlations in training data.

Dynamic adaptive attention mechanisms have recently been proposed as a way to modulate attention weights based on input uncertainty, thereby reducing the influence of unreliable tokens [17]. This line of work integrates supervised contrastive learning to pull representations of semantically similar examples together while pushing apart examples with different sentiment labels. By combining dynamic attention with contrastive objectives, the model can learn to discount noisy spans and focus on stable sentiment cues. Empirical results suggest that such hybrid approaches improve robustness under informal and perturbed conditions without sacrificing clean accuracy to the same degree as conventional adversarial training. The main structural trade-off involves computational overhead and the need for additional negative sampling strategies.

Interpretability methods can reveal whether a model relies on robust semantic signals or brittle surface correlations. Local explanation techniques assign importance scores to input features and can expose cases where a model bases its prediction on punctuation, hashtags, or spurious demographic markers [15]. Such explanations are valuable not only for debugging but also for auditing fairness and adversarial susceptibility. If an explanation shows that a sentiment classifier consistently attends to identity terms rather than evaluative language, the system likely encodes biases that can be triggered adversarially. Interpretability thus becomes a governance instrument rather than a purely technical add-on.

The architectural search space is shaped by a fundamental trade-off between representational capacity and robustness. Larger models with more parameters can memorize rare but stable sentiment patterns and may generalize better to diverse informal registers. At the same time, larger models tend to be more vulnerable to adversarial examples because their decision boundaries are more complex and can be approximated by attackers [16]. This trade-off implies that increasing model scale alone is not a robust deployment strategy. It must be accompanied by explicit robustness constraints, uncertainty estimation, and continuous monitoring.

## 4. Data Governance and Annotation Infrastructures

Robustness depends as much on data as on architecture. Most sentiment classification corpora are drawn from product reviews, movie reviews, or social media posts that have been heavily filtered. Such filtering removes exactly the noise and nonstandard forms that cause failures in deployment. Data governance must therefore include deliberate collection of noisy, informal, and dialectally diverse text to ensure that training distributions approximate operational distributions. Annotation guidelines must specify how to handle mixed-language utterances, sarcasm, emoji, and context-dependent sentiment. Without such specification, annotators fall back on inconsistent heuristics, and label noise compounds the robustness problem.

Provenance tracking is essential for auditing data lineage and identifying systematic gaps. A corpus that overrepresents English-language product reviews will not support robustness claims for multilingual social media sentiment. Data sheets and model cards have been proposed as institutional mechanisms for documenting training data, intended use cases, and known limitations [23]. These instruments create accountability and enable downstream users to assess whether a sentiment classifier is appropriate for their deployment context. They also facilitate the comparative evaluation of robustness interventions across research groups.

Label noise is a particularly insidious form of corruption because it degrades robustness even when input text is clean. In contrast to input perturbations, label errors are not directly observable from the model input and can persist through training without being detected. Supervised contrastive learning can partially mitigate label noise by structuring the representation space in a way that reduces reliance on individual noisy labels [17]. However, such methods cannot compensate for systematic annotation biases, such as the tendency to label dialectal expressions as more negative or less intelligent. Addressing those biases requires institutional changes in annotator recruitment, compensation, and quality control.

The construction of adversarial evaluation suites raises governance questions of its own. Attack frameworks such as TextAttack provide reproducible adversarial benchmarks and support a wide range of attack types [8]. While valuable for research, these benchmarks can create incentives to optimize for specific attack algorithms rather than for general robustness. A governance approach should require evaluation across multiple independent attack families, including character-level, word-level, phrase-level, and human-crafted adversarial examples. It should also include naturally occurring noisy text drawn from different platforms and communities. Only such multi-layered evaluation can reveal whether robustness improvements are genuine or narrowly overfit.

Data augmentation with synthetic misspellings, paraphrases, and Unicode confusables is widely used to improve robustness. Yet augmentation must be designed with careful attention to distributional realism. Randomly inserting typos into well-formed sentences does not reproduce the structured variation of informal language, such as phonetic spellings, eye dialect, or code-switching. Augmentation policies should therefore be informed by empirical studies of real-world text. They must also be updated as language changes, because slang, abbreviations, and adversarial strategies evolve over time. Static augmentation schedules become obsolete quickly, particularly in online environments where adversarial actors adapt rapidly.

## 5. Training Strategies and Dynamic Robustness Mechanisms

Training for robust sentiment classification requires moving beyond empirical risk minimization over clean data. Adversarial training, in which the model is exposed to perturbed examples during optimization, has been a dominant paradigm for improving worst-case robustness [18]. In text, adversarial training can be implemented by generating perturbations on the fly using gradient information or rule-based transformations [9]. This approach stabilizes decision boundaries and reduces sensitivity to small input changes. However, adversarial training introduces a tension between clean accuracy and robust accuracy. Models hardened against strong attacks often become conservative, assigning neutral predictions to ambiguous or informal examples that a clean model would classify confidently.

A complementary strategy is to use contrastive learning, which shapes the representation space directly rather than relying solely on classification loss. Supervised contrastive learning pulls together examples with the same sentiment label while pushing apart examples with different labels, even when their surface forms differ substantially [17]. When combined with dynamic adaptive attention, this approach enables the model to downweight tokens that are likely to be noisy or adversarial. The attention mechanism can learn to focus on stable sentiment spans, such as evaluative adjectives or verbs, while ignoring distractors introduced by attackers. This combination addresses a key weakness of standard attention models, which often attend to spurious but salient tokens.

Another robust training strategy involves data augmentation with adversarial examples generated by a separate attack model. Instead of solving a min-max optimization at every training step, the system can periodically generate a diverse set of attacks and incorporate them into the training corpus. This reduces computational cost and allows the model to learn from a wider range of perturbations. The risk is that the attack model becomes a single point of failure: if it fails to cover an important attack family, the trained classifier will remain vulnerable to that family. An ecosystem approach that rotates multiple attack models or uses human-in-the-loop red teaming can mitigate this risk.

Robustness to informal language also benefits from domain-adaptive pretraining. Continuing pretraining on a corpus of social media text before fine-tuning for sentiment classification can reduce the distributional mismatch between pretraining data and deployment data. However, domain-adaptive pretraining requires large quantities of unlabeled text that reflect the operational domain. It also raises privacy and copyright governance issues, particularly when the corpus includes user-generated content that cannot be redistributed. Institutional arrangements for data sharing and differential privacy must therefore be integrated into the training pipeline.

Dynamic inference-time defenses attempt to detect and neutralize adversarial inputs before classification. These methods can flag suspicious inputs based on prediction confidence, representation consistency, or external knowledge. For example, a system might compare the model output with and without stopword removal or character normalization. If the prediction changes dramatically, the input may be adversarial. Such defenses can reduce attack success rates but introduce additional latency and may themselves be attacked by adaptive adversaries who optimize inputs to evade detection. The deployment system must therefore include monitoring to identify when an inference-time defense is being probed and to update it accordingly.

## 6. Evaluation, Fairness, and Deployment Implications

Standard evaluation metrics such as accuracy and macro F1 are insufficient for robust sentiment classification. They aggregate performance over clean test sets and obscure differences in error rates across demographic groups, dialects, and perturbation types. Robustness evaluation must report disaggregated performance on noisy, informal, and adversarial subsets. It must also measure worst-case performance under a defined threat model, since average-case metrics can be misleading when an attacker can choose the input. Threat modeling should specify the attacker's knowledge, capabilities, and objectives. Without such specification, robustness claims are not operationally meaningful [19].

Fairness emerges as a central concern because robustness interventions can have disparate impacts. Adversarial training with synthetic misspellings may improve overall robustness

while degrading accuracy on dialectal text that shares surface similarities with those misspellings. Character-level models that are robust to typographical errors may inadvertently erase meaningful variation in African American English, code-switched text, or creole varieties. These risks have been extensively discussed in the broader natural language processing fairness literature [21]. The core issue is that robustness and fairness are not independent objectives; they interact through representation choices, data selection, and training objectives.

Language technology operates within power structures that shape whose language is considered standard and whose is treated as noise [22]. A robust sentiment classifier that normalizes nonstandard spelling to standard forms may systematically distort the sentiment expressed by minority language users. This is not merely a technical error but a form of symbolic erasure. Fairness audits should therefore examine the impact of robustness interventions on different language communities, not only on aggregate performance. Such audits require collaboration with linguists, community organizations, and domain experts who can interpret evaluation results in social context.

Deployment monitoring is essential for maintaining robustness after initial release. Language changes, new attack techniques emerge, and user populations shift. A system that was robust at deployment time may become vulnerable within months. Continuous evaluation requires collecting a representative sample of production inputs, labeling them with high-quality human annotations, and measuring drift. This is expensive and raises privacy concerns, but it is necessary for sustainable robustness. Monitoring must also include adversarial threat detection to identify coordinated attacks before they cause widespread harm. Incident response protocols should specify how to update models, filter inputs, or suspend service during active attacks.

Model governance for robust sentiment classification should include documentation requirements that extend beyond traditional model cards. Documentation should specify the threat model, the augmentation strategies, the evaluation datasets, the known failure modes, and the demographic fairness results. It should also record the provenance of training data and any third-party pretrained components. Such documentation supports external auditing and helps downstream users assess whether the system is suitable for their context. It also creates institutional memory that persists as team members change and systems are updated.

## 7. System-Level Trade-offs and Sustainability

Robustness improvements are rarely free. They consume additional computational resources, increase annotation costs, complicate training pipelines, and may reduce clean accuracy. A system design must therefore make explicit trade-offs among competing objectives. Decision makers must determine how much clean accuracy they are willing to sacrifice for improved adversarial robustness, how much latency they can tolerate for inference-time defenses, and how much annotation budget they can allocate to noisy and informal data. These decisions cannot be made by machine learning engineers alone; they require input from product managers, legal counsel, and affected communities.

The environmental sustainability of robustness interventions is an often overlooked dimension. Large-scale adversarial training and dynamic attention mechanisms require additional computation, which increases energy consumption and carbon footprint. Foundation models pretrained on massive corpora already carry substantial environmental costs [20]. Robustness enhancements that multiply these costs may undermine sustainability

commitments. System designers should therefore prefer computationally efficient robustness methods, such as targeted data augmentation or parameter-efficient fine-tuning, when they provide comparable benefits. Life-cycle assessment should be part of the governance framework for robust sentiment classification.

Maintainability is another sustainability concern. A robustness pipeline that depends on a complex stack of attack generators, contrastive objectives, and dynamic attention modules is difficult to maintain and debug. If a vulnerability emerges, engineers may struggle to isolate the responsible component. Modular design can mitigate this by separating the data augmentation layer, the representation learning layer, and the evaluation layer. Each module should have clear interfaces and version control, allowing independent updates without destabilizing the entire system. Institutional knowledge must be captured in documentation and automated tests so that robustness properties do not disappear during routine maintenance.

Cross-domain comparison reveals that robust sentiment classification shares structural challenges with other security-sensitive natural language processing tasks. Question answering, hate speech detection, and machine translation all face adversarial inputs and distributional shift. Lessons from these domains suggest that no single defense is sufficient. Instead, layered defenses that combine data hygiene, robust representation learning, adversarial training, inference-time checks, and continuous monitoring are necessary [19]. The same principle applies to sentiment classification. A system that relies solely on a pretrained model without monitoring will fail eventually, no matter how accurate it is on a static benchmark.

The economics of robustness also shape deployment decisions. In low-stakes applications such as social media analytics for trend detection, occasional misclassification may be acceptable, and the cost of adversarial training may not be justified. In high-stakes applications such as financial sentiment analysis for trading systems, patient feedback in healthcare, or public opinion monitoring for crisis response, misclassification can have severe consequences. Robustness investments should therefore be proportional to the operational risk. Governance frameworks can support risk-based decision making by providing clear taxonomies of failure consequences and threat likelihood.

## 8. Policy, Governance, and Future Directions

Policy approaches to robust sentiment classification must balance innovation, security, and civil rights. Regulatory regimes increasingly require algorithmic accountability for automated decision systems that affect individuals. Sentiment classification may be embedded in such systems, for example when customer service interactions influence credit decisions or when employee sentiment monitoring affects workplace evaluations. Robustness failures in these contexts can produce tangible harm. Policy should therefore require robustness testing under realistic distributional shift before deployment, as well as ongoing monitoring and incident reporting. These requirements should be proportionate to the risk level of the application.

International coordination is needed because adversarial attacks on text systems are not constrained by national borders. A sentiment classifier deployed by a multinational platform may be attacked from any jurisdiction, using adversarial examples generated with open-source tools. Shared threat intelligence and coordinated disclosure practices can help organizations respond quickly to emerging attack techniques. However, such coordination must respect privacy and avoid creating surveillance infrastructures that chill legitimate

expression. Policy frameworks should encourage transparency without mandating the disclosure of proprietary model internals unless necessary for public safety.

The future of robust sentiment classification will likely involve closer coupling between data governance and model training. Rather than treating data cleaning as a preprocessing step, organizations will integrate noise and variation into the training objective directly. Dynamic attention and contrastive learning are early examples of this trend [17]. Future systems may incorporate uncertainty quantification at the token level, enabling them to request clarification for ambiguous inputs or to abstain from classification when confidence is low. Such abstention mechanisms are particularly important for sentiment analysis, where context, sarcasm, and cultural references can make human judgment difficult even for annotators.

Human-in-the-loop governance will remain essential. Automated robustness testing can identify many vulnerabilities, but it cannot capture the full range of social and cultural meanings embedded in informal language. Human reviewers with deep knowledge of specific communities should be involved in evaluating the impact of robustness interventions on those communities. Their feedback can guide the design of augmentation policies, the selection of evaluation data, and the interpretation of fairness metrics. This is not a one-time consultation but an ongoing relationship that shapes system evolution.

Research directions should include the development of robustness benchmarks that span multiple languages, dialects, and platforms. Current benchmarks are heavily skewed toward English and toward a small number of adversarial attack families [8]. Expanding this coverage will require collaboration with linguists and community organizations. Research should also investigate the long-term dynamics of adversarial coevolution, in which attackers adapt to defenses and defenses adapt to attackers. This requires longitudinal studies that are rare in current literature. Finally, research should examine the environmental costs of robustness methods and develop efficiency metrics that go beyond accuracy.

## 9. Conclusion

Robust sentiment classification under noisy, informal, and adversarial text conditions is not a single technical problem but a system-level challenge that cuts across data governance, model architecture, training strategy, evaluation, deployment, and policy. This paper has argued that robustness emerges from interactions among these layers and cannot be achieved through isolated fixes. Character-level models, attention mechanisms, adversarial training, and contrastive learning each address part of the problem, but they also introduce new vulnerabilities and trade-offs. Dynamic adaptive attention combined with supervised contrastive learning represents a promising direction because it directly targets the problem of unreliable tokens and unstable representations [17]. However, even the most advanced architecture will fail without adequate data that reflects real-world noisy and informal language.

Fairness and sustainability must be treated as first-order constraints rather than afterthoughts. Robustness interventions can systematically disadvantage minority language communities if they are designed without attention to dialectal diversity and social context. Evaluation must therefore be disaggregated across demographic and linguistic groups, and documentation must support external audit. Computational costs must be weighed against environmental obligations. Governance frameworks should require proportionate robustness testing and continuous monitoring, particularly for high-stakes applications.

The path forward requires interdisciplinary collaboration among machine learning researchers, linguists, social scientists, legal scholars, and affected communities. It also requires a shift from static benchmark thinking to dynamic, context-aware system design. By embracing this complexity, the field can develop sentiment classification systems that are not only accurate on clean test sets but also reliable, equitable, and accountable in the messy and contested environments where they must operate.

**References**


1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

2. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186.

4. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. 3rd International Conference on Learning Representations.

5. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 31-36.

6. Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., & Chang, K.-W. (2018). Generating natural language adversarial examples. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2890-2896.

7. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(5), 8018-8025.

8. Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 119-126.

9. Miyato, T., Dai, A. M., & Goodfellow, I. (2017). Adversarial training methods for semi-supervised text classification. 5th International Conference on Learning Representations.

10. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, 28, 649-657.

11. Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746-1751.

12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

13. Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. 8th International Conference on Learning Representations.

14. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? China National Conference on Chinese Computational Linguistics, 194-206.

15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

16. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. 5th International Conference on Learning Representations.

17. Li, Q. (2026). Dynamic Adaptive Attention and Supervised Contrastive Learning: A Novel Hybrid Framework for Text Sentiment Classification. arXiv preprint arXiv:2604.10459.

18. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. 6th International Conference on Learning Representations.

19. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. IEEE European Symposium on Security and Privacy, 372-387.

20. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

21. Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 67-73.

22. Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of bias in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454-5476.

23. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

24. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15262-15271.

25. Pruthi, D., Dhingra, B., & Livingstone, A. (2019). Combating adversarial misspellings with robust word recognition. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5582-5591.