Systematic Generation of Adversarial Datasets with Controllable Noise Levels

Mohammad Reza Norouzi[1*]

[1] M.Sc. Student, Department of Computer Engineering, Kish International Campus, University of Tehran, Kish, Iran

*** Corresponding author email address**: mrnorouzi@ut.ac.ir

Received: 2025-10-01 **Revised:** 2026-02-07 **Accepted:** 2026-02-14 **Initial Publish:** 2026-04-25 **Final Publish:** 2026-09-01

Abstract

The rapid proliferation of user-generated textual content on social networks and digital platforms has created significant challenges for sentiment analysis systems. These challenges are more pronounced in the Persian language due to the scarcity of high-quality datasets, orthographic variability, and the high sensitivity of models to noise. One of the most critical issues is the vulnerability of machine learning models to textual noise and adversarial attacks, which can lead to substantial performance degradation. The objective of this study is to propose a systematic approach for generating adversarial textual datasets with controllable noise levels in order to evaluate and enhance the robustness of Persian sentiment analysis models. In this research, a baseline Persian sentiment analysis dataset was first preprocessed. Subsequently, a framework was designed to introduce targeted noise types, including word substitution, deletion, insertion, and permutation. For each type of noise, an intensity parameter was defined to enable precise control over the degree of perturbation. The adversarial data were generated independently of any specific model, and each instance was annotated not only with its sentiment label but also with metadata specifying the type and level of noise applied. The performance of several sentiment analysis models was then evaluated before and after training with the adversarial dataset. The results indicated that models trained exclusively on clean data experienced significant performance degradation when exposed to adversarial samples, particularly under substitution and deletion noise. In contrast, training with the generated adversarial dataset led to a considerable improvement in noise robustness and performance stability. The findings suggest that the systematic generation of adversarial data with controllable noise constitutes an effective instrument for sensitivity analysis and robustness enhancement in Persian sentiment analysis models and can play a critical role in the development of reliable systems under real-world conditions.

Keywords: *Persian sentiment analysis, adversarial data, controllable noise, natural language processing, model robustness.*

How to cite this article:

Norouzi, M. R. (2026). Systematic Generation of Adversarial Datasets with Controllable Noise Levels. Management Strategies and Engineering Sciences, 8(5), 1-9.

1. Introduction

The exponential growth of user-generated textual content across social media platforms, online marketplaces, and digital communication channels has fundamentally transformed the landscape of organizational decision-making and strategic management. Sentiment analysis, as a core task in natural language processing (NLP), has emerged as a critical analytical instrument for extracting opinions, attitudes, and emotional orientations from large-scale textual corpora [1, 2]. In management contexts, sentiment-driven insights support brand monitoring, customer experience optimization, risk assessment, and strategic forecasting [3, 4]. With the proliferation of big data infrastructures and deep learning architectures, sentiment analysis has shifted from rule-based and classical machine learning paradigms toward data-intensive, representation-learning approaches capable of capturing contextual and semantic nuances [5, 6]. However, despite substantial progress, contemporary sentiment analysis systems remain vulnerable to noise, domain shifts, and adversarial perturbations, raising concerns about their robustness and reliability in real-world managerial applications [7, 8].

The evolution of deep neural architectures has significantly enhanced text classification performance. Early neural models demonstrated that distributed representations

and compositional embeddings could rival syntactic feature engineering in text classification tasks [9, 10]. Convolutional and recurrent architectures, including LSTM and hybrid variants such as Co-LSTM, further improved contextual modeling and sequential representation learning for sentiment analysis in large-scale social data [11, 12]. With the introduction of the attention mechanism and transformer-based architectures, the field underwent a paradigm shift [13, 14]. Transformer models, particularly BERT and its derivatives, achieved state-of-the-art results in numerous NLP tasks, including sentiment classification [15, 16]. Comparative analyses have consistently demonstrated the superiority of transformer-based architectures over traditional recurrent or convolutional models in sentiment-related tasks [17, 18].

The Persian language, however, presents unique structural, morphological, and cultural complexities that complicate sentiment analysis. Feature engineering challenges, limited annotated datasets, orthographic variability, and tokenization issues remain significant barriers [19, 20]. Cultural and linguistic factors further influence polarity interpretation and pragmatic meaning, necessitating language-specific modeling strategies [21]. While transformer-based models such as ParsBERT and FaBERT have been developed to address Persian language understanding [22, 23], challenges persist in information retrieval, domain generalization, and standardized evaluation [24]. Even foundational preprocessing components such as stemming exhibit limitations in Persian due to morphological richness and dialectal diversity [25].

In practical management environments, sentiment analysis systems operate under noisy and uncontrolled conditions. Social media texts frequently contain misspellings, code-switching, informal language, and typographical errors [26, 27]. Persian sentiment analysis in e-commerce platforms has demonstrated sensitivity to preprocessing pipelines and embedding strategies [28]. Moreover, the increasing use of handwritten and electronic textual documents in digital ecosystems introduces additional heterogeneity in sentiment-bearing content [29]. These realities expose a critical gap between laboratory-level performance and real-world deployment robustness.

Beyond natural noise, adversarial attacks have emerged as a significant threat to NLP systems. Research has demonstrated that minimal perturbations—such as character swaps, synonym substitutions, or visually similar token manipulations—can drastically degrade model performance [30, 31]. Black-box adversarial generation techniques have successfully evaded deep learning classifiers by producing semantically similar yet adversarially crafted sequences [32]. TextBugger illustrated how adversarial text can deceive real-world applications through character-level and word-level modifications [33]. Even public toxicity detection systems have been shown to be vulnerable to subtle perturbations [34]. These findings underscore the fragility of sentiment models under adversarial conditions and highlight the necessity of robustness-oriented design.

Data augmentation and adversarial training have been proposed as potential mitigation strategies. Techniques such as easy data augmentation (EDA) introduced controlled lexical perturbations to improve generalization [35]. Back-translation has been employed at scale to generate semantically equivalent paraphrases that enhance model resilience [36]. Code-switching generation through adversarial networks has expanded cross-lingual robustness [37]. Adversarial training methods for large neural language models have demonstrated improved resistance to perturbations [38]. In parallel, robustness research in NLP has provided comprehensive taxonomies of adversarial defenses and evaluation frameworks [7]. However, systematic and controllable adversarial dataset construction remains underexplored, particularly in low-resource languages such as Persian.

Optimization and model stability further complicate robustness efforts. Hyperparameter tuning plays a decisive role in deep learning performance [39]. Frameworks such as Optuna enable Bayesian optimization and pruning strategies for efficient parameter search [40]. Gradient accumulation techniques and modular training pipelines enhance scalability and reproducibility in large language model training [41, 42]. Nevertheless, improvements in optimization do not inherently guarantee robustness against structured adversarial noise.

In management-driven applications, the reliability of sentiment analytics directly influences strategic outcomes. Big data frameworks integrating sentiment classification into decision-support systems require stable and interpretable outputs [3, 43]. Cognitive-inspired analytics architectures have emphasized resilience and explainability in large-scale sentiment systems [4]. Real-time sentiment monitoring for institutional or organizational analysis necessitates robustness under dynamic and noisy input streams [26]. When sentiment models are susceptible to adversarial perturbations or uncontrolled textual distortions, managerial decisions derived from such systems risk being biased or unreliable.

The theoretical underpinnings of transformer interpretability and circuit-level analysis further suggest that understanding internal model behavior is crucial for robustness enhancement [42]. Representation erasure techniques have revealed how specific components contribute to classification decisions [10]. Cross-lingual transformer performance has shown surprising transferability but also potential vulnerability across languages [44]. These insights indicate that robustness must be addressed both at the data level and architectural level.

Although extensive surveys have documented advances and challenges in NLP-based sentiment analysis [6, 8], and adversarial defense research has expanded considerably [7], the intersection of systematic adversarial dataset generation, controllable noise modeling, and Persian sentiment analysis remains insufficiently investigated. Existing Persian sentiment research has focused largely on feature engineering, embeddings, and architecture comparison [19, 20], while adversarial robustness has received comparatively limited attention. Furthermore, real-world sentiment systems often rely on heterogeneous transformer architectures, including ParsBERT, FaBERT, and multilingual BERT variants [15, 22, 23], yet cross-architecture robustness validation under systematically generated adversarial conditions has rarely been conducted.

Given the critical managerial implications of sentiment-driven analytics, the increasing sophistication of adversarial attacks, and the linguistic particularities of Persian, there is a compelling need for a structured and controllable framework that generates adversarial textual datasets capable of evaluating and strengthening model robustness across architectures and noise levels [7, 32, 33]. Such a framework must integrate insights from data augmentation [35], adversarial training [38], transformer optimization [40], and Persian language modeling challenges [21, 25].

Accordingly, the aim of this study is to design and empirically validate a systematic framework for generating Persian adversarial textual datasets with controllable noise levels in order to evaluate and enhance the robustness of transformer-based sentiment analysis models in real-world management applications.

2. Methodology

This study is applied–experimental in nature and was conducted with the objective of designing and implementing a systematic framework for generating adversarial textual datasets with controllable noise levels in the domain of Persian sentiment analysis. The research procedure comprised four principal stages: preparation of the baseline dataset, design of the noise model, generation of adversarial data, and validation of the generated dataset.

In the first stage, a baseline dataset was selected and preprocessed. For this purpose, a validated Persian sentiment analysis corpus consisting of labeled texts (positive, negative, and neutral) was utilized. Data preprocessing included normalization of Persian characters, removal of unnecessary characters, standardization of spacing, and text tokenization. The purpose of this stage was to ensure data uniformity prior to the introduction of noise and to prevent interference between preprocessing errors and the targeted perturbations.

In the second stage, a systematic noise model was designed. Within this model, noise was categorized into four principal types:

1. Lexical substitution (using synonyms, colloquial expressions, or words with different sentiment polarity);
2. Lexical deletion (removal of key or non-key words);
3. Insertion of irrelevant or neutral words;
4. Reordering of words or phrases.

For each type of noise, a noise intensity parameter was defined, representing the percentage of manipulated words relative to the length of the text. This parameter enabled the generation of datasets with low, medium, and high noise levels and provided precise control over the degree of semantic distortion introduced into the text.

In the third stage, adversarial data were generated automatically and independently of any specific model. For each textual instance in the baseline dataset, multiple adversarial versions were produced using different combinations of noise types and intensity levels. In this process, the original sentiment label of each text was preserved, and supplementary metadata specifying the type of noise and the applied intensity level were recorded. This labeling structure enables precise analysis of the impact of each noise type on model performance.

In the final stage, the generated dataset was validated. To this end, several commonly used sentiment analysis models trained on clean data were evaluated on the adversarial dataset and subsequently compared with their performance after being trained using the proposed generated dataset. The results demonstrated that employing the generated adversarial dataset led to a statistically significant improvement in model robustness against noise and

adversarial attacks. This validation process confirms the effectiveness and practical applicability of the proposed approach.

3. Findings and Results

Noise Rate Configuration

To enable the examination of different noise levels, each adversarial task was controlled by a specific percentage that determined the proportion of that task within the final dataset. In addition to sampling rates, a maximum edit constraint was embedded for each transformation to ensure that the number of substitutions, insertions, or deletions in each sentence remained limited and controllable. Consequently, texts shorter than a predefined minimum threshold were excluded from certain tasks to prevent excessive distortion of very brief inputs. For example, in the sentence truncation task, if a sentence contained fewer than three words, the task was not applied. This constraint was imposed to avoid complete destruction of the sentence meaning. For instance, if a sentence consisted of only one word (e.g., “good,” “excellent,” “satisfied”), deletion would entirely eliminate its semantic content.

Adversarial Tasks

Single-Task Transformations

To achieve maximum coverage of perturbations present in large-scale textual datasets, 14 primary adversarial data generation tasks in Persian were employed as follows:

Synonym Substitution: One or more words were replaced with their Persian synonyms without altering the sentence meaning. Example: “I am satisfied with my purchase.” → “I am happy with my purchase.”

Random Word Insertion: Irrelevant words without semantic relation were added to test the model’s ability to ignore noise. Example: “I am satisfied with my purchase.” → “I am satisfied flower with my purchase.”

Stopword Insertion: High-frequency, low-information words (e.g., “that,” “from,” “for”) were inserted to increase syntactic complexity. Example: “I am satisfied with my purchase.” → “I am from satisfied with my purchase.”

Spelling Error Injection: Characters were modified or misplaced to simulate typographical errors. Example: “I am satisfied with my purchase.” → “I am satistfied with my purchase.”

Character Swap: Two adjacent characters within a word were swapped, creating a common typing error. Example: “I am satisfied with my purchase.” → “I am satifsied with my purchase.”

Character Deletion: A character was removed from a word, producing an incomplete or erroneous form. Example: “I am satisfied with my purchase.” → “I am satisfed with my purchase.”

Keyboard Error Simulation: A character was replaced with a neighboring character on the Persian keyboard. Example: “I am satisfied with my purchase.” → “I am satisfjed with my purchase.”

Extra Character Insertion: An additional unrelated character was inserted within a word, altering its visual structure. Example: “I am satisfied with my purchase.” → “I am satisfiied with my purchase.”

Word Order Permutation: The order of words was modified in a controlled manner while preserving all tokens. Example: “I am satisfied with my purchase.” → “Satisfied I am with my purchase.”

Sentence Truncation: Part of the beginning or end of the sentence was removed, resulting in incomplete input. Example: “I am satisfied with my purchase.” → “I am satisfied with.”

Back Translation: The sentence was translated into another language and then translated back into Persian, producing structural variation while preserving semantic equivalence.

Noise Injection: Random characters, symbols, or meaningless words were inserted into the sentence. Example: “I am satisfied with my purchase.” → “I am satisfied & with my purchase.”

Code-Switching: A portion of the Persian sentence was replaced with an English word or phrase. Example: “I am satisfied with my purchase.” → “I am satisfied with my purchase.” (with one word replaced by its English equivalent).

Named Entity Replacement: Nominal entities, such as names of persons or places, were replaced with another entity of the same type. Example: “I am satisfied with my purchase.” → “I am satisfied with my product.”

Multi-Task Transformations

Because noise in large-scale datasets is neither predictable nor necessarily isolated, combinations of the above tasks were also applied to ensure that the model achieved maximum preparedness for adversarial datasets and remained robust under uncontrolled real-world scenarios. Accordingly, pairs of initial transformations were constructed to create more challenging perturbations. For example, applying typographical errors together with stopword insertion simultaneously tests lexical robustness, while combining noise injection with back translation

merges semantic paraphrasing with low-level character disruption. Controllable constraints and percentages were defined for these combinations as well. For each defined transformation (single- or multi-task), the number of instances to be modified was first calculated based on the configured percentage of the total dataset. The selected function was then applied to the designated number of rows, with edits performed up to the predefined limit. Finally, the modified texts, along with their original sentiment labels and the transformation name, were exported in a .csv file format.

Fine-Tuning on Adversarial Datasets and Comparative Evaluation

Initially, the base architecture was trained on noise-free data for one quarter of the total epochs. This stage enabled the model to learn core sentiment patterns under clean conditions. Subsequently, the model underwent three consecutive fine-tuning stages on increasingly challenging adversarial datasets (Levels 1, 2, and 3). By progressively exposing the network to mild perturbations and severe distortions, it adapted to diverse noise patterns. This curriculum learning strategy prevented the model from being overwhelmed by severe noise while simultaneously enhancing robustness against adversarial data. In practical terms, initial training on clean data established stable feature representations, and subsequent stages improved flexibility against controlled adversarial distortions. During each stage, the same tokenizer and gradient accumulation model were maintained to preserve stability. Optimizer settings, learning rate, and regularization parameters remained constant to ensure that performance differences were attributable solely to dataset difficulty. After completing all training stages, both the baseline model trained on normal data and the adversarially fine-tuned model were evaluated on an identical noisy dataset to enable precise performance comparison. To confirm the generalizability of the proposed adversarial fine-tuning strategy across architectures, the entire process was replicated using three pre-trained BERT variants: FaBERT, ParsBERT, and Multilingual BERT. This replication increased confidence in the applicability of the method and reduced the likelihood that the observed performance gains were architecture-specific.

Evaluation Metrics

To evaluate classification performance, commonly used and critical metrics were employed. For comprehensive comparison, Accuracy, Precision, Recall, F1-score, Area Under the Curve (AUC), and the Confusion Matrix were examined for both models.

Accuracy measures the proportion of correct predictions among all predictions; however, it may be misleading in imbalanced datasets:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision represents the proportion of true positive predictions among all positive predictions made by the model. High precision corresponds to a low false positive rate:

$$\text{Precision} = TP / (TP + FP)$$

Recall (Sensitivity) reflects the proportion of actual positive instances correctly identified by the model. High recall corresponds to a low false negative rate:

$$\text{Recall} = TP / (TP + FN)$$

F1-score is the harmonic mean of Precision and Recall and reflects the balance between them. It is particularly useful in scenarios involving class imbalance:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Area Under the Curve (AUC) refers to the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various threshold settings. A higher AUC indicates better discriminative performance.

Standard Error of the Mean (SEM) quantitatively estimates the precision of the sample mean relative to the population mean. In model evaluation, SEM can be used to assess variability across different data samples or cross-validation folds, providing insight into the reliability of performance estimates.

Confusion Matrix

The confusion matrix is a tabular representation of actual versus predicted classifications, offering insight into the types of errors generated by the model. It displays true positives, false positives, true negatives, and false negatives, facilitating computation of performance metrics and highlighting specific areas where misclassification occurs.

Precision–Recall Curves

These curves plot precision against recall at different threshold settings, providing a comprehensive view of the trade-off between the two metrics. They are particularly informative in imbalanced classification scenarios where accuracy may be less meaningful.

Implementation

All experiments were conducted in a Google Colab Pro+ environment equipped with an NVIDIA A100 GPU featuring 40 GB of VRAM, a 12-core CPU, and 85 GB of RAM. The entire codebase was implemented in Python 3.9. Key libraries installed at runtime included custom modules for model construction and training loops, the Hugging Face

Transformers library for loading and fine-tuning multiple BERT variants, Optuna for Bayesian hyperparameter optimization and pruning, and additional utilities for dataset splitting and metric computation. This integrated setup ensured sufficient memory for gradient accumulation, accelerated attention computation on A100 tensor cores, and full reproducibility of preprocessing, adversarial data generation, multi-stage training, and evaluation procedures within a shareable experimental framework.

4. Discussion and Conclusion

The findings of this study demonstrate that transformer-based sentiment analysis models trained exclusively on clean Persian datasets experience statistically significant performance degradation when exposed to systematically generated adversarial noise. This result is consistent with prior research indicating that neural NLP models are highly sensitive to minor lexical, character-level, and structural perturbations [30, 33]. Similar vulnerabilities have been documented in black-box adversarial settings, where semantically preserved yet perturbed inputs effectively deceive deep learning classifiers [32]. The substantial drop observed in substitution, deletion, and character-level noise conditions aligns with findings that visually or orthographically altered tokens disrupt embedding representations and downstream classification decisions [31, 45]. In particular, character swaps and misspellings compromise tokenization and subword segmentation mechanisms, a phenomenon previously highlighted in transformer robustness analyses [46].

The degradation patterns observed across BERT-based architectures reinforce the argument that contextual embeddings, despite their superior representational power, remain susceptible to structured perturbations. Although transformer models outperform recurrent and convolutional baselines under clean conditions [17, 18], their internal attention mechanisms do not inherently guarantee robustness against adversarial manipulations [7]. The decline in performance across ParsBERT, FaBERT, and Multilingual BERT indicates that architectural sophistication alone is insufficient to counteract adversarial distortions. This observation is consistent with broader surveys emphasizing that state-of-the-art NLP systems often prioritize accuracy over resilience [6, 8].

The Persian-specific results further underscore linguistic and morphological factors that amplify adversarial impact. Persian exhibits rich morphology, flexible word order, and orthographic variability, all of which can intensify the consequences of minor perturbations [19, 20]. Cultural nuances embedded in sentiment expressions may also be distorted by synonym substitution or code-switching transformations [21]. The observed sensitivity to stemming-related variations supports prior evidence that Persian morphological processing remains imperfect, even with advanced NLP pipelines [25]. Consequently, adversarial perturbations in Persian can propagate through tokenization and embedding stages more severely than in languages with more standardized orthographic conventions.

Importantly, the study reveals that adversarial fine-tuning significantly mitigates performance degradation. Models exposed to progressively increasing noise levels demonstrated improved robustness across all evaluation metrics, including Accuracy, Precision, Recall, F1-score, and AUC. This improvement aligns with adversarial training literature suggesting that exposure to perturbed samples enhances model generalization and defensive capacity [7, 38]. The gradual fine-tuning strategy implemented in this study resembles curriculum-based learning paradigms, where models adapt incrementally to challenging inputs rather than being overwhelmed by severe distortions at early stages. Such progressive adaptation supports findings that structured data augmentation improves resilience without sacrificing baseline performance [35, 36].

The effectiveness of controlled multi-task adversarial combinations further highlights the importance of realistic perturbation modeling. Real-world textual data rarely contain isolated noise types; rather, they exhibit layered distortions including typos, informal language, and semantic shifts [26, 27]. By simulating compound perturbations, the framework approximated authentic noisy environments, thereby enhancing external validity. Prior studies have emphasized that combining perturbation strategies increases the robustness of classification systems compared to single-task augmentation [37]. The present findings extend this insight to the Persian language domain and demonstrate cross-architecture generalizability.

The replication of the adversarial fine-tuning process across ParsBERT, FaBERT, and Multilingual BERT reinforces the robustness of the proposed framework. Transformer-based Persian models have previously shown strong baseline performance in language understanding tasks [22, 23]. However, comparative reviews indicate that transformer variants may exhibit heterogeneous performance patterns depending on dataset characteristics and preprocessing strategies [15, 24]. The consistent

robustness improvements observed across architectures suggest that the proposed adversarial dataset generation strategy functions independently of specific model design choices.

The managerial implications of these findings are significant. Sentiment analysis systems are increasingly integrated into big data decision-support infrastructures [3, 43]. Cognitive-inspired analytical frameworks emphasize reliability and adaptability in large-scale environments [4]. If sentiment classifiers remain vulnerable to adversarial or noisy input, managerial decisions derived from these systems may be biased or strategically misleading. The demonstrated performance stabilization achieved through adversarial fine-tuning directly contributes to enhancing the trustworthiness of sentiment-driven analytics in organizational contexts.

From an optimization perspective, the controlled experimental setup ensured that robustness gains were attributable to dataset manipulation rather than hyperparameter variation. Hyperparameter optimization frameworks such as Optuna have been shown to influence deep learning performance significantly [39, 40]. By maintaining consistent optimization parameters and employing stable gradient accumulation strategies [41, 42], the study isolates adversarial exposure as the principal explanatory variable. This methodological rigor strengthens the causal inference that systematic adversarial data generation enhances model robustness.

The results also contribute to theoretical discussions on model interpretability and representation stability. Prior work using representation erasure demonstrated that neural classifiers rely on fragile lexical cues [10]. The observed improvements after adversarial training suggest that exposure to controlled perturbations encourages models to develop more distributed and semantically grounded representations. This aligns with earlier demonstrations that deep unordered composition can rival syntactic methods when models are appropriately trained [9]. Furthermore, the resilience improvements parallel findings in other modalities, where adversarial examples have been shown to enhance recognition robustness when incorporated into training [47].

In summary, the study confirms that Persian transformer-based sentiment models are highly vulnerable to adversarial perturbations under clean training regimes, but that systematic, controllable adversarial dataset generation combined with progressive fine-tuning substantially enhances robustness, generalizability, and managerial reliability. These findings contribute to both applied NLP research and management analytics by bridging the gap between high-performance laboratory models and resilient real-world deployment.

Despite its contributions, this study is subject to several limitations. First, the adversarial transformations were rule-based and predefined, potentially limiting coverage of more sophisticated semantic-level attacks. Second, although multiple transformer architectures were evaluated, the experiments focused primarily on BERT-based models and did not include alternative generative large language models. Third, the evaluation was conducted on a specific Persian sentiment dataset, and generalization to other domains, dialects, or multimodal sentiment contexts remains uncertain. Finally, computational constraints may have restricted exploration of extremely large-scale datasets or ultra-deep architectures.

Future research may explore adaptive adversarial generation using reinforcement learning or generative adversarial networks to simulate more complex semantic manipulations. Expanding the framework to cross-domain and cross-lingual datasets would further test generalizability. Investigating interpretability-aware adversarial defenses could clarify how representation changes contribute to robustness. Additionally, integrating large language models and instruction-tuned architectures may reveal new robustness dynamics. Finally, exploring the interaction between adversarial robustness and fairness metrics would provide deeper insight into ethical and managerial implications.

From a practical perspective, organizations employing Persian sentiment analysis systems should incorporate adversarial robustness evaluation into their model validation pipelines. Training strategies should progressively expose models to realistic noise patterns to enhance deployment stability. Monitoring performance across diverse noise levels can improve risk management in sentiment-driven decision support systems. Furthermore, maintaining consistent optimization settings and systematic dataset documentation can strengthen reproducibility and accountability in managerial analytics infrastructures.

Authors' Contributions

Authors equally contributed to this article.

Acknowledgments

Authors thank all participants who participate in this study.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

References

[1] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731-5780, 2022.

[2] A. Saxena, H. Reddy, and P. Saxena, "Introduction to sentiment analysis covering basics, tools, evaluation metrics, challenges, and applications," in *Principles of social networking: the new horizon and emerging challenges*, 2022, pp. 249-277.

[3] A. P. Rodrigues and N. N. Chiplunkar, "A new big data approach for topic classification and sentiment analysis of Twitter data," *Evolutionary Intelligence*, pp. 1-11, 2022.

[4] D. K. Jain, P. Boyapati, J. Venkatesh, and M. Prakash, "An intelligent cognitive-inspired computing with big data analytics framework for sentiment analysis and classification," *Information Processing & Management*, vol. 59, no. 1, p. 102758, 2022.

[5] K. Chowdhary and K. R. Chowdhary, "Natural language processing," in *Fundamentals of artificial intelligence*, 2020, pp. 603-649.

[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713-3744, 2023.

[7] Y. Zhou, D. Jin, and X. Ren, "A Survey of Adversarial Defenses and Robustness in Natural Language Processing," *arXiv preprint*, 2022.

[8] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 100059, 2024.

[9] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé Iii, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

[10] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," *arXiv preprint*, 2016.

[11] R. K. Behera, M. Jena, S. K. Rath, and S. Misra, "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data," *Information Processing & Management*, vol. 58, no. 1, p. 102435, 2021.

[12] A. Berrajaa, "Natural language processing for the analysis sentiment using a LSTM model," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022, doi: 10.14569/IJACSA.2022.0130589.

[13] A. Vaswani and et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] B. Ghojogh and A. Ghodsi, "Attention mechanism, transformers, BERT, and GPT: tutorial and survey," 2020, doi: 10.31219/osf.io/m6gcn.

[15] M. V. Koroteev, "BERT: a review of applications in natural language processing and understanding," *arXiv preprint*, 2021.

[16] D. Rothman, *Transformers for Natural Language Processing*. Packt Publishing Ltd., 2022.

[17] K. Pipalia, R. Bhadja, and M. Shukla, "Comparative analysis of different transformer based architectures used in sentiment analysis," in *2020 9th international conference system modeling and advancement in research trends (SMART)*, 2020: IEEE, pp. 411-415.

[18] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," *Knowledge and Information Systems*, vol. 66, no. 12, pp. 7305-7361, 2024.

[19] R. Asgarnezhad and S. A. Monadjemi, "Persian sentiment analysis: feature engineering, datasets, and challenges," *Journal of applied intelligent systems & information sciences*, vol. 2, no. 2, pp. 1-21, 2021.

[20] Z. Rajabi and M. Valavi, "A survey on sentiment analysis in Persian: a comprehensive system perspective covering challenges and advances in resources and methods," *Cognitive Computation*, vol. 13, no. 4, pp. 882-902, 2021.

[21] R. Shokrzad, "The Impact of Culture on Persian NLP: A Linguistic Perspective," *Journal of Language and AI Ethics*, 2023.

[22] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding," *Neural Processing Letters*, vol. 53, pp. 3831-3847, 2021.

[23] M. Masumi, S. S. Majd, M. Shamsfard, and H. Beigy, "FaBERT: Pre-training BERT on Persian Blogs," ed, 2024.

[24] S. Moniri, T. Schlosser, and D. Kowerko, "Investigating the Challenges and Opportunities in Persian Language Information Retrieval through Standardized Data Collections and Deep Learning," *Computers*, vol. 13, no. 8, p. 212, 2024.

[25] M. Assadi, V. Shaghaghi, and M. Kahani, "Capabilities and Limitations of Persian Stemming in Natural Language Processing," *Research in Western Iranian Languages and Dialects*, vol. 13, no. 1, pp. 1-17, 2025.

[26] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Real-time Twitter Sentiment Analysis for Moroccan Universities using Machine Learning and Big Data Technologies," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 5, 2023.

[27] D. Paulraj, P. Ezhumalai, and M. Prakash, "A Deep Learning Modified Neural Network (DLMNN) based proficient sentiment analysis technique on Twitter data," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 36, no. 3, 2024.

[28] S. Shumaly, M. Yazdinejad, and Y. Guo, "Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fastText embeddings," *PeerJ Computer Science*, vol. 7, p. e422, 2021.

[29] R. Ahamad and K. N. Mishra, "Exploring sentiment analysis in handwritten and E-text documents using advanced machine learning techniques: a novel approach," *Journal of Big Data*, vol. 12, no. 1, p. 11, 2025.

[30] Y. Belinkov and Y. Bisk, "Synthetic and natural noise both break neural machine translation," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
[31] S. Eger and D. Benz, "Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems," *arXiv preprint*, 2020, doi: 10.18653/v1/N19-1165.
[32] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW)*, 2018.
[33] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "TextBugger: Generating adversarial text against real-world applications," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2019.
[34] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving Google's Perspective API built for detecting toxic comments," *arXiv preprint*, 2017.
[35] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
[36] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, doi: 10.18653/v1/D18-1045.
[37] C. H. Chang and Y. C. Lin, "Code-switching sentence generation by generative adversarial networks and its application to data augmentation," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.
[38] X. Liu and et al., "Adversarial training for large neural language models," *arXiv preprint*, 2020.
[39] A. Aghaebrahimian and M. Cieliebak, "Hyperparameter tuning for deep learning in natural language processing," in *4th swiss text analytics conference (swisstext 2019)*, 2019: SwissText.
[40] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623-2631, doi: 10.1145/3292500.3330701.
[41] J. Lamy-Poirier, "Layered gradient accumulation and modular pipeline parallelism: fast and efficient training of large language models," *arXiv preprint*, 2021.
[42] A. Wang and D. Xiao, "Understanding how LLMs complete a classical NLP task by gradient accumulation-based circuit discovery," in *Third International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2024)*, 2024, vol. 13181: SPIE, pp. 859-866.
[43] A. Alarifi and et al., "A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks," *The Journal of Supercomputing*, vol. 76, pp. 4414-4429, 2020.
[44] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," *arXiv preprint*, 2019, doi: 10.18653/v1/D19-1077.
[45] D. Pruthi, B. Dhingra, and Z. C. Lipton, "Combating adversarial misspellings with robust word recognition," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
[46] S. J. Mielke *et al.*, "Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP," *arXiv preprint*, 2021.
[47] C. Xie and et al., "Adversarial examples improve image recognition," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.