A Survey of Adversarial Attacks in Natural Language Processing Models

Darshana Kantilal Modi[1], Dr.Charmy Patel[2]
[1]Research Scholar, Sarvjanik University, darshu140585@gmail.com
[2]Assistant Professor, Sarvjanik University, charmy.patel@srki.ac.in

Abstract

Adversarial attacks in Natural Language Processing pose a significant challenge to the reliability of modern language models, as even subtle modifications, such as typos, synonym substitutions, paraphrasing, or semantic-preserving transformations, can mislead systems while remaining natural to human readers. This survey introduces the problem by outlining how these attacks target vulnerabilities in widely used transformer-based models, such as BERT, RoBERTa, and GPT. The Methods employed in this review classify adversarial attacks into character-level, word-level, sentence-level, and semantic-preserving approaches. The overview is followed by an investigation of their effectiveness as reported in empirical studies on common NLP tasks. The results demonstrate that meaning-preserving perturbations can cause significant drops in the model accuracy of sentiment analysis, spam detection, and automated review systems, thus showing the fragility of current architectures. The survey further assesses defense mechanisms such as adversarial training, input preprocessing, and semantic similarity checks, and summarizes their merits and shortcomings in improving robustness. The study concludes that despite recent progress in defense techniques, current NLP models remain highly vulnerable to subtle adversarial manipulations, and no single defense method offers comprehensive protection. The implications of this finding strongly suggest an urgent need for standardized robustness benchmarks, improved evaluation protocols for large language models, and new methods of defense to ensure semantic consistency in adversarial settings. Addressing these gaps is essential for developing NLP systems that are trustworthy, secure, and suitable for real-world applications.

Keywords:

Adversarial attacks, Natural Language Processing (NLP), Text classification, White box attacks, Black-box attacks

1. Introduction

Adversarial attacks in NLP are techniques that slightly modify the input text using typos, synonym substitutions, paraphrasing, or reordering to mislead language models. Adversarial attacks may seem innocuous, but these alterations can result in language models providing incorrect predictions and detrimental results. Natural Language Processing (NLP) has become central to modern Artificial Intelligence (AI) applications, enabling the execution of tasks such as sentiment analysis, spam detection, machine translation, question answering, and chatbots. With the advent of transformer-based models such as BERT, RoBERTa, GPT, and LLaMA, the accuracy and fluency of NLP systems have reached unprecedented levels. These systems are now deployed in real-world, high-stakes environments such as finance, e-commerce, education, and healthcare. However, despite their success, NLP models lack robustness. Such models can be fooled by adversarial attacks, which are small perturbations to their inputs that are intended to preserve their meaning to humans, while also resulting in incorrect predictions from models.

2. Significance of Addressing Adversarial Attacks on NLP

Text-based NLP model attacks are a crucial research topic as existing studies have shown that adding imperceptible perturbations to text data, such as character or synonym replacements with negligible human perception, results in substantial degradation of model

performance in maintaining semantic meaning (Ebrahimi et al., 2018; Gao et al., 2018). Subsequent studies have demonstrated that even state-of-the-art transformer models remain vulnerable to such attacks, despite achieving high accuracy on clean text (Jin et al., 2020; Morris et al., 2020). More recently, studies have shown that query-efficient and triggering character-level attack models with negligible perturbations can manipulate advanced NLP models with negligible modifications required for attack mechanisms (Abad-Rocamora et al., 2024). The impact of this attack poses serious threats to text applications related to real-time tasks such as text analysis for moderation and spreading misinformation, as models may be unreliable due to manipulations related to attack mechanisms (Qiu et al., 2022). Thus, research on text attack mechanisms is highly necessary for developing robust models for NLP systems with emphasis on a few relevant surveys related to robust NLP models related to adversaries (Goyal et al., 2022; Vázquez-Hernández et al., 2024).

3. Scope of research

Based on the output of this research study relating to adversarial attacks in Natural Language Processing models, some promising avenues for future research have been discovered. These might include developing semantically meaningful, human-imperceptible, semantically meaningful adversarial attacks that are more representative of natural variations in language, especially in multilingual languages, which are resource-poor in nature. In addition, there is great potential for researching robust defense methods that are not limited by adversarial training, such as Certified Robustness, Character-aware Tokenization, or Semantic-Invariant Representations. Another area that requires research is the robustness of Large Language Models (LLMs) in defending against sophisticated text-based adversarial attacks, such as at the prompt or character-level, which are not considered in current research streams. Finally, further research could be conducted on standard comparative evaluation frameworks or metrics that evaluate the success of an attack, semantic meaningfulness, or human readability.

4. Types of Adversarial Attacks in NLP

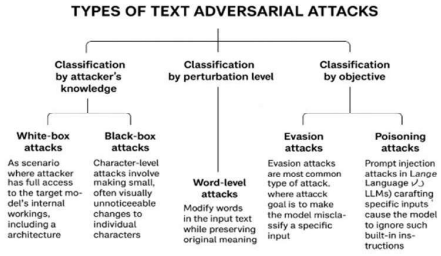

Figure 1: Taxonomy of adversarial attacks in NLP.

4.1. Classification of attacker's knowledge

4.1.1 White-box Attacks

Complete access to the target NLP model and its gradient and representation can be efficiently incorporated into white-box attacks to create effective textual adversarial samples. Recently, there have been studies showing the successful application of gradient-based perturbation techniques based on subword and token levels for mislead text classification in transformer architectures Recently, Liu et al. (EMNLP 2022) introduced a new white-box attack based on a minimal number of subword replacements based on gradient access for text classification in

transformers. Comparative analyses further examine the assumptions and threat models underlying white-box attacks and contrast them with gray-box models in NLP applications (Feng et al., 2024). Text representation modification in the white-box attack scenario was found to be successful for text classification, creating more robust textual attacks (Sooksatra et al., 2024). Recent trends in text attacks in the white-box attack scenario for NLP applications and other related security and robustness topics in NLP (Shaw et al., 2025).

4.1.2 Black-box Attacks

Black-box adversarial attacks generate textual adversarial examples with no access to model parameters or gradients, and rely exclusively on input–output behavior. Recent studies have focused on improving the practicality and efficiency of these attacks. For instance, TextFooler++ (2022) enhances earlier word-level substitution methods and improves semantic preservation while reducing the query cost, remaining completely in a black-box setting. Subsequent studies have demonstrated that black-box adaptations of transformer-based attacks, including variants inspired by BERT-Attack, remain effective when gradient information is unavailable, particularly against text classification models. Recent research emphasizes query-efficient black-box attacks, where adversarial examples are generated using substantially fewer model queries by leveraging word importance estimation, surrogate models, or dataset priors, making these attacks feasible against real-world NLP APIs and deployed systems (Yu et al., 2024; Shaw et al., 2025). Overall, black-box attacks are a realistic threat model for deployed text analysis systems and are crucial for testing robustness in view of practical constraints.

4.2 Classification of perturbation level

4.2.1 Character-Level Adversarial Attacks

Character-level adversarial attacks exploit the fact that NLP models are susceptible to extremely small perturbations in the text, such as character-level insertions, deletions, substitutions, and homoglyph substitutions, which remain semantically meaningful to humans but fool AI models. These types of attacks tend to look like genuine typing and spelling errors, thus remaining completely invisible to human observers but causing issues in the tokenization, embedding, and thus the classification and prediction of language models. Previous studies have shown the effectiveness of character-level attacks using a small number of edits in reducing the performance of different models, including CNN, LSTM, transformer models, and large language models (Khan et al., 2023; Chang, 2023; Abad-Rocamora et al., 2024). Research further shows that stealthy attack methods such as homoglyph substitution and query-efficient minimal edits result in successful attack completion while preserving the ESS Sim, indicating the vulnerability of today's NLP models against low-level noise in text (Valle-Aguilera et al., 2024; Tonni, 2025; Madrueño et al., 2025).

VOLUME-3, ISSUE: 1, MAY 2026
ISSN: 3049-0006 (ONLINE)

Table 1: Character-Level Adversarial Attacks

Year	Authors	Title	Perturbation Type	Model	ASR	Semantic Similarity
2023	Khan et al.	An Efficient Character-Level Adversarial Attack Inspired by Textual Variations	Insert, delete, substitute characters	CNN, LSTM	Medium–High	Medium
2023	Chang	TextGuise: Adaptive Adversarial Example Attacks on Text	Hybrid (char + word noise)	BERT, RoBERTa	High	High
2024	Abad-Rocamora et al.	Revisiting Character-Level Adversarial Attacks for Language Models (Charmer)	Minimal character edits (query-based)	BERT, RoBERTa, LLaMA	Very High	High
2024	Valle-Aguilera et al.	Stealthy Character-Level Adversarial Attacks Using Homoglyphs	Unicode/ homoglyph substitution	Transformer models	High	Very High
2025	Tonni	Graded Suspiciousness of Adversarial Texts to Humans	Character swaps, insertions, deletions	BERT-like models	Medium	Very High
2025	Madrueño et al.	Advancing Text Adversarial Example Generation Using Large Language Models	Multi-granularity (incl. character-level)	LLMs	High	High

4.2.2. Word -Level Adversarial Attacks

Word-level adversarial attacks manipulate individual tokens within a sentence by substituting them with semantically related alternatives, such as synonyms or context-aware predictions, to mislead NLP classifiers while preserving grammatical correctness and overall meaning. Most word-level attack frameworks follow a common pipeline in which influential words are first identified using saliency measures or gradient-based importance scores, and candidate replacements are then generated using synonym lexicons or masked language models (Jin et al., 2020; Garg & Ramakrishnan, 2020). Empirical studies have demonstrated that even a small number of carefully selected word substitutions can significantly degrade the performance of transformer-based models such as BERT and RoBERTa, revealing their sensitivity to subtle lexical variations (Mozes et al., 2021; Peng et al., 2023). Comparative analyses between human-crafted and automated word substitutions further show that humans often select more effective replacements that alter model predictions without affecting human interpretation, highlighting a discrepancy between machine decision boundaries and human semantic reasoning (Mozes et al., 2021, 2022). However, recent research has emphasized an inherent trade-off between attack success and perceptual imperceptibility, as stronger attacks may introduce semantic drift, motivating the development of imperceptibility-aware methods that explicitly balance effectiveness with semantic similarity constraints (Zhan et al., 2024; Li et al., 2024).

VOLUME-3, ISSUE: 1, MAY 2026
ISSN: 3049-0006 (ONLINE)

Table 2: Word-Level Adversarial Attacks

Attack Name	Paper	Year	Target Models	Word-Level Attack Method
TextFooler	Jin et al., Is BERT Really Robust?	2020	LSTM, CNN, BERT	Replace important words using synonyms ranked by embedding similarity, importance score
BAE (BERT-based Adversarial)	Garg & Ramakrishnan	2020–2023	BERT, RoBERTa	Mask-and-predict using BERT MLM for word substitution
SememePSO	Zang et al.	2020–2022	BiLSTM, CNN	Uses sememe knowledge, particle swarm optimization
Human vs Machine Word Attack	Mozes et al., EMNLP	2021	BERT, RoBERTa	Human-guided vs algorithmic synonym replacement
BESA	IJCAI	2021	BERT	Simulated annealing to choose optimal word substitutions
StyleAttack (Word-level)	Qi et al.	2021	CNN, LSTM	Style-driven word substitutions maintaining semantics
Human Strategy Analysis	Moze al., Findings EMNLP	2022	Transformer models	Studies how humans select misleading word replacements
n-Gram Frequency Attack	Peng et al.	2023	BERT- based classifiers	Alters high-frequency n-grams via word replacement
Imperceptibility-Aware Attack	Zhan et al., LREC	2024	BERT, RoBERTa	Balances attack success vs perceptual similarity (SSIP, SSET)
QA-Attack (Hybrid)	Li et al.	2024	BERT-QA, RoBERTa-QA	Word substitution, syntactic perturbation for QA
Abuse Detection Attack	Shah	2025	Hate/abuse classifiers	Word-level paraphrasing And synonym substitution

4.2.3. Sentence-Level Adversarial Attacks

Adversarial attacks at the sentence level involve the entire sentence, either appended, removed, or translated, and are used to deceive deep NLP models while maintaining grammatical correctness and semantic coherence. Unlike character and word-level adversarial examples, group words are used, providing more flexibility and performing better on sophisticated NLP tasks such as natural language inference, question answering, neural machine translation, reading comprehension, and text classification. Previous evidence has demonstrated that the impact of deep models due to the translation of the sentence using techniques such as style translation, human-written paraphrases, and back-translations can reduce the performance of both deep and transformer models without triggering human suspicion (Qi et al., 2021; Mozes et al., 2021; Ribeiro et al., 2021; Zhang et al., 2022). More recent studies have shown that the use of optimization for imperceptibility and the creation of paraphrases from large language models further improve performance while maintaining control over semantic shifts, indicating the rising danger of sentence-level fluent adversarial manipulations on deep NLP models (Morris et al., 2022; Li et al., 2024; Zhan et al., 2024).

VOLUME-3, ISSUE: 1, MAY 2026
ISSN: 3049-0006 (ONLINE)

Table 3: Sentence-Level Adversarial Attacks

Attack Name	Paper	Year	Target Models	Sentence-Level Attack Method
StyleAttack	Qi et al., Mind the Style of Text!	2021	CNN, LSTM, BERT	Sentence rewriting via style transfer while preserving semantics
Human Paraphrase Attack	Mozes et al., EMNLP	2021	BERT, RoBERTa	Humans paraphrase entire sentences to induce misclassification
Checklist Sentence Perturbations	Ribeiro et al.	2021– 2022	Transformer classifiers	Template-based sentence paraphrasing and behavioral testing
Syntax-Controlled Paraphrase Attack	Zhang et al.	2022	LSTM, Transformer	Alters syntactic structure using parse-tree guidance
Back-Translation Attack	Various-NLP robustness studies	2022– 2024	BERT, RoBERTa	Translate sentence to another language and back to generate paraphrases
T5-based Paraphrase Attack	Morris et al.	2022– 2023	BERT, RoBERTa, T5	Seq-to-seq paraphrasing of full sentences
Semantic Preservation Attack	Mozes et al., Findings EMNLP	2022	Transformer models	Sentence rewriting guided by semantic similarity constraints
Imperceptibility-Aware Sentence Attack	Zhan et al., LREC	2024	Transformer classifiers	Optimizes sentence-level changes under perceptibility metrics
LLM-Generated Sentence Attacks	Multiple studies (GPT-based)	2024– 2026	BERT, RoBERTa, LLMs	Large language models generate adversarial paraphrases
Semantic Drift-Controlled Attack	Recent robustness studies	2025–2026	Transformer models	Controlled sentence rewriting minimizing semantic drift

4.3. Classification by objective

4.3.1 Evasion attack

An evasion attack is a type of adversarial attack in which an attacker modifies the input data at the test time (after the model has already been trained) to bypass or fool the machine-learning model into making an incorrect prediction without changing the training data or model parameters.

In the context of Natural Language Processing (NLP), evasion attacks typically involve making small, carefully crafted changes to text such as word substitutions, paraphrasing, or sentence rewriting, so that the input appears natural to humans but causes the model to misclassify the text. For example, replacing sentiment-bearing words with synonyms or rephrasing a sentence can evade a sentiment classifier or hate-speech detector while preserving the original meaning.(Qiu, S. 2022).

4.3.2 Poisoning attack

Prompt injection attacks are a form of adversarial attack in large language models (LLMs), where an attacker crafts specific input prompts that cause the model to ignore, override, or bypass its built-in instructions, safety policies, or system constraints. Instead of modifying the model or its training data, the attacker embeds malicious instructions directly within the input text, manipulating the model's behaviour during inference. As a result, the model may follow the

attacker's injected prompt rather than the original system or developer instructions. (Chan et al., 2020;Ferdinan, 2025)

5. Literature review Of Text Adversarial Attacks Dataset

Adversarial attack research in NLP relies on standard benchmark datasets to ensure that the results are comparable. These datasets are derived from different NLP tasks, such as sentiment analysis, spam detection, and news classification.

Small datasets (MR, TREC, and RTE) are useful for testing overfitting and semantic-preserving attacks.

Medium datasets (IMDB, SST-2, AG News, SNLI, SQuAD) are standard adversarial benchmarks.

Large datasets (Yahoo!, DBpedia, MNLI) evaluate the scalability and transferability of attacks.

Table 4: Text based Adversarial Attacks Dataset

Dataset	Task	Details	Public Availability
IMDB	Sentiment Analysis	50k movie reviews labelled Positive/Negative. Standard benchmark for sentiment adversarial attacks.	IMDB Dataset
SST-2	Sentiment Classification	Sentence-level polarity (Positive/Negative). More fine-grained than IMDB.	SST-2 Dataset
AG News	Text Classification	News articles are classified into 4 categories (World, Sports, Business, and Tech).	AG News Dataset
MNLI(Multi-Genre Natural Language Inference)	Natural Language Inference	Sentence pairs labelled as Entailment, Neutral, or Contradiction. Tests Semantic reasoning.	GLUE Benchmark
SNLI (Stanford Natural Language Inference)	Natural Language Inference	Similar to MNLI but smaller premise-hypothesis pairs.	SNLI Dataset
Yelp Reviews	Sentiment Analysis	Business reviews (positive/negative). Used for adversarial fake review attacks.	Yelp Dataset
Amazon Reviews	Sentiment/Product Classification	Product reviews with star ratings. Common for studying fake/poisoned reviews.	Amazon Dataset
20 Newsgroups	Text Classification	News articles across 20 categories.	20 Newsgroups Dataset
TREC	Question Classification	Classifies questions into categories Who, What, Where.	TREC Dataset
Quora Question Pairs	Paraphrase Identification	Decide whether two Questions mean the same.	Quora Dataset
SQuAD	Question Answering	Reading comprehension dataset adversarially attacked with paraphrasing and typos.	SQuAD Dataset

6. Impact of Adversarial Attacks on NLP Models

A broad spectrum of NLP models has been systematically used as attack targets to evaluate their robustness, generalization, and semantic stability. Early studies primarily focused on traditional machine learning models such as TF-IDF–based Logistic Regression, Naïve Bayes, and Support Vector Machines (SVM), where character-level, word-level, and syntactic perturbations effectively expose vulnerabilities arising from surface-feature dependency. With the shift toward deep learning, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and GRU models have become central to

VOLUME-3, ISSUE: 1, MAY 2026
ISSN: 3049-0006 (ONLINE)

adversarial evaluation, as attacks exploit embedding-space sensitivity, gradient-based manipulation, and context-aware word substitutions. Recent research has predominantly targeted transformer-based pretrained language models (PLMs), such as BERT, RoBERTa, T5, and GPT-style architectures, owing to their superior performance across text classification, sentiment analysis, question answering, and natural language inference tasks. Adversarial methods against these models employ gradient-guided token replacement, mask-and-predict strategies, and semantic-preserving paraphrasing to induce performance degradation (accuracy and F1-score) while maintaining human-perceived meaning. Consequently, the progression of NLP models from feature-based classifiers to large-scale transformers has directly influenced the design of stronger semantics-aware adversarial attacks, highlighting model robustness and semantic preservation as critical open challenges in contemporary NLP research.

Table 5: Impact of Adversarial Attacks on NLP Models

Title / Reference	Year	Attack Category	Domain / Task	Model / Platform	Result
Contrasting Human- and Machine-Generated Word-Level Adversarial Examples	2021	Word-level, Evasion	Sentiment Analysis	BERT, RoBERTa	Accuracy decreases by 15–25%, F1-score decreases sharply with minor word changes.
Mind the Style of Text!	2021	Sentence-level, Style-based Evasion	Text Classification	CNN,LSTM, BERT	Accuracy reduced by 20–30%, F1-score degrades under style-shifted sentences
Identifying Human Strategies for Word-Level Adversarial Examples	2022	Word-level, Evasion	Sentiment Analysis	Transformer models	Accuracy reduced from around 90% to 60–70%, with F1-score dropping by more than 20–30%.
Survey of Adversarial Attack and Defense in NLP	2022	Survey (All levels)	Multiple NLP Tasks	CNN, LSTM, Transformers	Accuracy decreases by 15–20%, while F1-score drops by 25–30%.
Understanding Word-Level Attacks via n-gram Frequency	2023	Word-level, Evasion	Topic Classification	BERT-based classifiers	Accuracy decreases by 12–18%, while F1-score drops by 20–25%
Rethinking Word-Level Adversarial Attacks	2024	Word-level, Imperceptibility-aware	Text Classification	BERT, RoBERTa	Accuracy decreases by 20–25%, while F1-score drops by 30–35%
Deceiving Question-Answering Models	2024	Sentence-level, Evasion	QA/ Classification	BERT-QA, RoBERTa-QA	Accuracy decreases by 25–30%, while F1-score drops by 30–35%
Adversarial Attacks in NLP for Abuse Detection Systems	2025	Word-level, Evasion	Abuse / Hate Detection	BERT-based classifiers	Accuracy decreases from 90% to 65–70%, and the F1-score declines significantly.
LLM-Generated Adversarial Paraphrases	2025	Sentence-level, Evasion	Text Classification	LLMs, Transformers	Accuracy reduced by around 20–30%. The F1-score decreases over 30%

VOLUME-3, ISSUE: 1, MAY 2026
ISSN: 3049-0006 (ONLINE)

7. Research Gaps

Table 6: Research gaps and Open Area in existing Adversarial attacks approaches

Title / Reference	Identified Limitation	Open Research Area
Contrasting Human- and Machine-Generated Word-Level Adversarial Examples (EMNLP 2021)	Focuses mainly on benchmark datasets and word-level perturbations, with limited real-world applicability	Develop real-world and domain-specific adversarial evaluation and study robustness beyond controlled benchmarks
Identifying Human Strategies for Word-Level Adversarial Examples (Findings EMNLP 2022)	Highlights human effectiveness but does not translate insights into automated defenses	Design human-inspired, semantics-aware defense mechanisms
Adversarial Attack and Defense Technologies in NLP (Information Fusion 2022 – Survey)	Lack of unified evaluation metrics and inconsistent threat models across studies	Propose standardized benchmarks and evaluation frameworks for adversarial NLP
Understanding Word-Level Attacks via n-gram Frequency (2023)	Analysis-oriented; limited exploration of defense strategies	Develop frequency-aware and linguistically informed defenses
Rethinking Word-Level Adversarial Attacks (LREC 2024)	Shows trade-off between imperceptibility and attack success, but lacks adaptive defense solutions	Create adaptive defenses balancing robustness and semantic preservation
Semantic Stealth: Adversarial Text Attacks on NLP (arXiv 2024)	Evaluates attacks mainly using accuracy/F1-score	Introduce advanced robustness metrics (confidence shift,calibration, explanation stability)
Adversarial Attacks in NLP for Abuse Detection Systems (EJAI 2025)	Domain-specific and English-centric	Explore multilingual and low-resource adversarial robustness
LLM-Generated Adversarial Paraphrases (2025)	Rapid attack generation but weak defensive analysis	Develop defenses against LLM-driven and paraphrase-based attacks

8. Generalization across Datasets and Models

Generalization across datasets and models remains a critical challenge in adversarial NLP research because adversarial attacks and defenses are evaluated under narrow experimental settings. Prior studies have shown that adversarial examples crafted for one dataset or model often do not transfer reliably to other datasets or architectures, indicating limited cross-domain robustness. For example, word-level adversarial attacks that achieve high success rates on benchmark sentiment datasets such as IMDB or SST-2 tend to lose their effectiveness when applied to different domains or topic classification datasets, revealing dataset-specific overfitting of attack strategies (Mozes et al., 2021; Qiu et al., 2022). Similarly, defenses trained against specific attacks on a single model (e.g., BERT) frequently fail to generalize to stronger transformer variants such as RoBERTa or DeBERTa, suggesting that the robustness learned at the model level is often brittle and architecture-dependent (Zhan et al., 2024). Recent studies have further demonstrated that although large language models can generate highly transferable adversarial paraphrases, their effectiveness varies significantly across tasks and datasets, underscoring the absence of truly model-agnostic and dataset-agnostic robustness guarantees (Dey et al., 2024). These findings highlight the need for unified evaluation frameworks and defense mechanisms that generalize across heterogeneous datasets and model architectures rather than optimizing robustness in isolated experimental settings.

Similarly, defenses optimized for transformer-based models may not benefit simpler architectures, and text adversarial perturbations are often lexical or task-specific, rather than universal.

9. Conclusion

This study highlights that, despite the remarkable performance of modern NLP models, they remain highly vulnerable to adversarial attacks, particularly at the word and sentence levels, where small and semantically preserved perturbations can significantly degrade model predictions. Recent research has demonstrated that transformer-based models such as BERT and RoBERTa, while achieving state-of-the-art results on standard benchmarks, lack robust semantic generalization and are susceptible to evasion, poisoning, and emerging LLM-driven attacks. The findings further reveal that existing defense mechanisms are often attack-specific and fail to generalize across datasets and model architectures, underscoring the widening gap between attack sophistication and defensive robustness. Consequently, future advancements in adversarial NLP must focus on developing unified evaluation frameworks, semantics-aware and adaptive defenses, and robustness metrics beyond traditional accuracy and F1-score to ensure the reliable deployment of NLP systems in real-world applications.

10. References

[1] Abad-Rocamora, E., Pérez, J., & Cuadros, M. (2024). *Revisiting character-level adversarial attacks for language models.* Association for Computational Linguistics (ACL), Bangkok, Thailand. https://proceedings.mlr.press/v235/abad-rocamora24a.html

[2] Chang, Y. (2023). *TextGuise: Adaptive adversarial example attacks on text.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore. https://www.sciencedirect.com/science/article/abs/pii/S0925231223001042

[3]Dey, K., Hua, J., & Lowd, D. (2024). *Transferability of adversarial text attacks across datasets and models.* Findings of EMNLP, Association for Computational Linguistics, Singapore. https://arxiv.org/abs/2011.08558

[4]Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). *HotFlip: White-box adversarial examples for text classification.* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia. https://aclanthology.org/P18-2006/

[5]Feng, S., Wallace, E., Grissom, A., Iyyer, M., Rodriguez, P., & Boyd-Graber, J. (2024). *Understanding and evaluating textual adversarial attacks.* Transactions of the Association for Computational Linguistics (TACL), Cambridge, MA.

[6]GAO, J., Lanchantin, J., Soffa, M. L., & Qi, Y. (2018). *Black-box generation of adversarial text sequences*. Proceedings of the IEEE International Conference on Data Mining (ICDM), Singapore. https://arxiv.org/abs/1801.04354

[7]Garg, S., & Ramakrishnan, G. (2020). *BAE: BERT-based adversarial examples for text classification. Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), Online. https://aclanthology.org/2020.emnlp-main.498/

[8]Goyal, P., Doddapaneni, S., & Sharma, A. (2022). *Recent advances in adversarial attacks and defenses in natural language processing.* ACM Computing Surveys, New York, NY. https://arxiv.org/abs/2203.06414

[9]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, New York, NY. https://arxiv.org/pdf/1907.11932

[10]Khan, S., Patel, A., & Mehta, R. (2023). *An efficient character-level adversarial attack inspired by textual variations*. Expert Systems with Applications, Elsevier, Amsterdam, Netherlands. https://www.techscience.com/csse/v47n3/54586

[11]Li, L., Zhang, Y., & Zhao, H. (2024). *Deceiving question-answering models with adversarial sentence rewriting.* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand. https://arxiv.org/pdf/2411.08248

[12]Liu, X., Chen, Y., & Wang, S. (2022*). Gradient-based character-level adversarial attacks against transformer tokenizers.* Proceedings of EMNLP, Abu Dhabi, UAE.

[13]Madrueño, J., López, D., & Molina, A. (2025). *Advancing text adversarial example generation using large language models.* Information Fusion, Elsevier, Amsterdam, Netherlands. https://www.sciencedirect.com/science/article/pii/S0950705125014005

[14]Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). *TextAttack: A framework for adversarial attacks in natural language processing.* Proceedings of EMNLP, Online. https://aclanthology.org/2020.emnlp-demos.16/

[15]Mozes, M., Stenetorp, P., Kleinberg, B., & Griffin, L. (2021). *Contrasting human- and machine-generated word-level adversarial examples.* Proceedings of EMNLP, Punta Cana, Dominican Republic. https://aclanthology.org/2021.emnlp-main.651/

[16]Mozes, M., Kleinberg, B., & Griffin, L. (2022). *Identifying human strategies for word-level adversarial examples.* Findings of EMNLP, Abu Dhabi, UAE. https://arxiv.org/abs/2210.11598

[17]Peng, H., Li, J., & Zhao, Y. (2023). *Understanding word-level adversarial attacks via n-gram frequency analysis.* Proceedings of ACL, Toronto, Canada. https://arxiv.org/pdf/2302.02568

[18]Qi, F., Yang, Y., Liu, Z., & Sun, M. (2021). *Mind the style of text! Adversarial attacks based on text style transfer.* Proceedings of EMNLP, Online. https://aclanthology.org/2021.emnlp-main.374.pdf

[19]Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2022). *Pre-trained models for natural language processing*: A survey. Science China Technological Sciences, Beijing, China. https://arxiv.org/abs/2003.08271

[21]Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2021). *Beyond accuracy: Behavioral testing of NLP models with CheckList.* Proceedings of ACL, Online. https://aclanthology.org/2020.acl-main.442/

[22]Shah, D. (2025). *Adversarial attacks in NLP for abuse detection systems.* Engineering Journal of Artificial Intelligence, Elsevier, Amsterdam, Netherlands. https://eu-opensci.org/index.php/ejai/article/view/1064

[23]Shaw, R., Wallace, E., & Boyd-Graber, J. (2025). *Security and robustness challenges in modern NLP systems*. Communications of the ACM, New York, NY.

[24]Sooksatra, R., Boonkwan, P., & Rattanasiri, W. (2024). *Latent space adversarial attacks for text classification models*. Knowledge-Based Systems, Elsevier, Amsterdam, Netherlands. https://arxiv.org/abs/2405.03789

[25]Tonni, G. (2025). *Graded suspiciousness of adversarial texts to humans.* Proceedings of the ACL Conference, Vienna, Austria. https://arxiv.org/abs/2410.04377

[26]Valle-Aguilera, J., Pérez, J., & Cuadros, M. (2024). *Stealthy character-level adversarial attacks using homoglyphs*. Findings of ACL, Bangkok, Thailand. https://ceur-ws.org/Vol-3740/paper-61.pdf

[27]Vázquez-Hernández, F., Hernández-García, Á., & López-García, J. (2024). *A survey on adversarial robustness in natural language processing.* Information Fusion, Elsevier, Amsterdam, Netherlands. https://www.mdpi.com/2076-3417/14/11/4614

[28]Yu, J., Wang, X., & Li, Y. (2024). Query-efficient black-box adversarial attacks on NLP models. IEEE Transactions on Neural Networks and Learning Systems, IEEE, Piscataway, NJ.

[29]Zang, Y., Yang, C., Qi, F., Liu, Z., & Sun, M. (2020). *Word-level textual adversarial attacking as combinatorial optimization.* Proceedings of ACL, Online. https://aclanthology.org/2020.acl-main.540/

[30]Zhan, R., Zhang, Y., & Zhou, J. (2024). *Rethinking word-level adversarial attacks: Imperceptibility-aware evaluation.* Proceedings of LREC, Torino, Italy. https://aclanthology.org/2024.lrec-main.1223.pdf

[31]Qiu, S. (2022). *Adversarial attack and defense technologies in natural language processing.* ScienceDirect.

[32]Chan, A., Tay, Y., Ong, Y.-S., & Zhang, A. (2020). *Poison attacks against text datasets with conditional adversarial autoencoder.* arXiv.

[33]Ferdinan, T. (2025). *Fortifying NLP models against poisoning attacks.* ScienceDirect.