Adversarial Natural Language Processing: Overview, Challenges and Future Directions

Laxmi Shaw
Texas State University
hde23@txstate.edu

Mohammed Wasim Ansari
Texas State University
ojq6@txstate.edu

Tahir Ekin
Texas State University
tahirekin@txstate.edu

Abstract

Natural language processing (NLP) has gained wider utilization with the emergence of large language models. However, adversarial attacks threaten their reliability. We present an overview of adversarial NLP with an emphasis on challenges, emerging areas and future directions. First, we review attack methods and evaluate the vulnerabilities of popular NLP models. Then, we review defense strategies including adversarial training. We identify key trends and suggest future directions such as the use of Bayesian methods to improve the security and robustness of NLP systems.

Keywords: NLP, Adversarial Machine Learning, Text Classification, AI Security, Bayesian Methods.

1. Introduction

Natural language processing (NLP) has undergone significant evolution over the decades. It transitioned from rule-based systems to machine learning and statistical models. Recently, the use of deep learning and the introduction of transformer architectures marked a revolutionary moment for NLP (Devlin et al. (2018)). This has redefined human-computer interaction and broadened the scope of NLP applications in various domains fueled by the emergence of generative AI (GenAI) and large language models (LLMs). Models such as BERT and generative pre-trained transformers (GPT) have yielded cutting-edge performance in tasks such as language understanding, generation, translation, and summarization (Johri et al. (2021)). NLP based technologies enhanced various fields such as healthcare, customer service, education, and entertainment (Esmradi et al. (2023)).

While there is great potential in use of NLP, increasing reliance on these systems amplifies the security concerns. Adversarial attacks can manipulate input data to impact NLP outcomes. For instance, adversarial attacks can alter the sentiment of a text, manipulate translation results, or generate misleading content in automated systems. The consequences of adversarial attacks on NLP systems can be severe and range from security and privacy risks to reduced reliability. They could reduce trust in NLP systems, especially in critical applications like legal document analysis, medical diagnosis, or autonomous vehicles. To mitigate the impact of adversarial threats, various defensive techniques such as adversarial training (AdvT) and input preprocessing are utilized (Goyal et al. (2023)). Incorporating adversarial testing and evaluation in the development process can help identify and fix potential weaknesses before deploying NLP systems in real-world applications. Despite these efforts, ensuring the security of NLP systems remains an ongoing battle, requiring continuous innovation and adaptation.

Several review papers have surveyed adversarial attacks and defenses in NLP focusing on different aspects. W. E. Zhang et al. (2020) provides a comprehensive overview of adversarial attack techniques on deep learning models in NLP, covering convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers. They emphasize the diversity of attack methods and their impact on various NLP tasks. X. Li et al. (2020) focuses on the vulnerability of RNNs to adversarial attacks, exploring how spatial and temporal dependencies in text data can be exploited, and highlighting potential mitigation strategies. Dong et al. (2022) covers both adversarial attacks and defenses on NLP in deep learning. Alsmadi et al. (2022) presents a survey of methods for text generation. Y. Cheng et al. (2019) examines the susceptibility of neural machine translation (NMT) models, particularly transformers, to adversarial perturbations. These reviews focus on specific models or types of attacks without providing a unified overview. In addition, recent advancements in adversarial attack methods and

URI: https://hdl.handle.net/10125/108946
978-0-9981331-8-8

defense mechanisms are not fully captured in these surveys, particularly empirical comparisons involving ensemble techniques developed in the last few years. These reviews also often often lack guidance on the practical implementation making it challenging for practitioners to apply these methods. In terms of LLM defenses, Esmradi et al. (2023) analyze LLM security vulnerabilities and review effective defense strategies including data sanitization, encryption-based methods, differential privacy and filtering. The review of Qiu et al. (2022) is the most similar to our manuscript in that they cover various attack methods, such as character-level, word-level, and sentence-level perturbations, and defense strategies, from data augmentation and AdvT to recent innovations like certified defenses, in addition to their discussion of evaluation metrics. While their coverage is more extensive compared to ours, it does not include recent advances, practical guidance and a discussion of emerging techniques such as Bayesian methods.

This manuscript reviews adversarial attacks in NLP, examining attack methods, exploited vulnerabilities, and defense strategies. We identify key trends, gaps, and future directions, with a focus on Bayesian methods. The contributions include a holistic overview that integrates recent attack and defense techniques, such as LLM attacks and zero shot defenses, practical guidance including some empirical comparisons to assess method effectiveness, and coverage of emerging techniques such as Bayesian methods.

This manuscript proceeds as follows. Section 2 provides an overview of the literature of NLP methods with an emphasis on Bayesian methods. Section 3 presents an overview of adversarial attacks, existing defenses and provides practical guidance. Section 4 presents emerging areas and future research directions. The manuscript concludes with final remarks in Section 5.

2. Related Literature

2.1. Overview of NLP Techniques

The main NLP methodologies include rule-based approaches, statistical methods, machine learning (ML), deep learning (DL), and transformer models. The choice of the algorithm depends on the application because of the varying strengths and weaknesses of these approaches. Rule-based approaches use predefined linguistic rules for tasks like tokenization and parsing. While they are interpretable and preferred for simple tasks, they may lack adaptability to natural language complexity. Statistical methods such as hidden Markov models (HMMs) and conditional random fields are utilized for tasks like part-of-speech (POS) tagging and named entity recognition (NER). They are robust and handle uncertainty. However, they require extensive feature engineering, and may not capture long-range dependencies. Supervised ML models, e.g., support vector machines and naive Bayes classifiers, learn from labeled data for tasks like text classification. Unsupervised ML methods, like clustering and topic modeling, uncover hidden structures in text. While ML algorithms may overfit to training data and need large labeled datasets and feature engineering, they are versatile and generalize well to new data. DL techniques, such as RNNs, CNNs, and long short-term memory networks (LSTMs) learn hierarchical representations from raw text. They excel at sequence modeling, text classification, and machine translation but require large datasets. The ability to capture complex patterns and dependencies in text may become computationally expensive possibly suffering from gradient issues. Transformer architectures like BERT and GPT use self-attention mechanisms to capture contextual relationships in text. They excel at language understanding, generation, translation, and summarization but need massive computational resources and extensive pre-training on large text corpora (Vaswani et al. (2017)). While they achieve state-of-the-art performance across NLP tasks, computing and data needs may limit their use.

2.2. Bayesian methods for NLP

Bayesian approaches are versatile for NLP applications due to their natural ability to model uncertainty, embed prior knowledge and decision making under incomplete information (Cohen (2022)). In NLP tasks such as text classification, sentiment analysis, or machine translation, Bayesian inference can be used to estimate the posterior distribution of model parameters given observed data. This allows for the incorporation of prior knowledge, which can help in regularizing models to avoid overfitting and improve generalization to unseen data. For instance, Naive Bayes classifiers, are foundational in text classification tasks. They are based on Bayes' theorem and make the simplifying assumption that the features are conditionally independent given the class label. Despite this simplification, Naive Bayes often performs surprisingly well, especially in spam detection, sentiment analysis, and topic classification.

For probabilistic clustering of text data, a Bayesian hierarchical method, Latent Dirichlet Allocation (LDA) (Blei et al. (2003)) and the subsequent development of topic models have become popular for document summarization and information retrieval (Abdelrazek et al. (2023)). Bayesian methods also can be used for sequence labeling tasks such as POS tagging. For example, HMMs, which are probabilistic models that assume a sequence of

observed words is generated by a sequence of hidden states (POS tags), can be trained using Bayesian inference. Bayesian networks and HMMs can be used to model the probabilistic relationships that arise in machine translation, word sense disambiguation, information retrieval, parsing and named entity recognition. Finally, LLMs where the goal is to predict the probability of a sequence of words benefit from Bayesian n-gram models and neural networks in capturing the probabilistic relationships between words (Chien (2019)). They can be useful for speech recognition and text generation.

While Bayesian methods provide robust and interpretable solutions to a wide range of NLP tasks, their applications in adversarial NLP is still limited. The computational complexity and cost of Bayesian models may be one of the reasons that prohibit applications for some large datasets and using complex models.

3. Adversarial NLP

In conducting the following literature survey, we have followed the guidelines listed by Webster and Watson (2002), and Brocke et al. (2009). Our coverage is deemed as a combination of exhaustive with selective citations and central focusing on select topics of adversarial attacks and defenses in natural language processing. In particular, we have used keywords "Natural language processing", "Adversarial attack NLP", "Adversarial defense NLP" for queries in "Google Scholar", "IEEE Explore", and "ScienceDirect". We have utilized a backward and forward search focusing on the attack and defense methods that are published between the years of 2018 and 2024. However, various synonyms for the term "natural language processing", such as "text mining" or "computational linguistics", have been disregarded in our study, highlighting its incomplete nature. In addition, we omitted the preprints that have less than 50 citations. Figure. 1 presents the taxonomy for the adversarial attack and defense methods in NLP.

3.1. Adversarial Attacks in NLP

Adversarial attacks with varying complexity and knowledge levels have been developed in order to weaken the performance and reliability of NLP systems. These methods usually involve changing text data, at different levels such as characters (char), words or sentences, to trick NLP models to produce incorrect outputs as in misclassification. Character-level attacks involve altering individual characters in a text to trick NLP models, such as changing "cat" to "c at". This minor modification can disrupt the model's understanding and cause incorrect predictions (Huang et al. (2021)). In word-level attacks, individual words in a text are strategically altered to mislead NLP models without changing the overall meaning. An example involves changing "The product is excellent" to "The product is terrible" by substituting "excellent" with "terrible". This subtle change can cause the model to misclassify the sentiment (Gao et al. (2018)). Sentence-level (paraphrasing) attacks rephrase sentences to confuse NLP models while keeping the meaning the same. An example is rephrasing "The cat sat on the mat" to "The mat was where the cat sat" that can lead the model to misclassify this sentence (Zeng et al. (2018)).

In terms of knowledge of the attacked model, adversarial attacks are at varying levels between white-box and black-box attacks. In white-box attacks, the attacker has complete information about the target model, including its architecture, parameters, and training data. A black-box attacker has little or no knowledge of the target model. They rely on querying the model and observing its outputs to create adversarial examples. Binary attacks corresponds to attacks for binary classifications. Attacks can also be targeted with a particular objective, or may be more general as non-targeted.

The attack methods broadly range from basic attacks with adding manually crafted inputs or or altering words, to iterative gradient based refinements. Our work reflects the advances in NLP models where attackers used heuristic and gradient based methods to exploit model vulnerabilities. Heuristic methods methods are mostly specific to certain models lacking generalizability. Gradient-based techniques like the Fast Gradient Sign Method (FGSM) create perturbations that misled models while remaining undetectable to humans (R. Wang (2022)). These gradient based adversarial perturbations could be iteratively refined, as in iterative FGSM and Projected Gradient Descent (PGD), resulting in higher success rates than one-shot methods (Chao et al. (2023)). Table 1 presents an overview of highlights of adversarial attacks in NLP ranging from char-level to sentence-level attacks.

3.2. Defense Mechanisms in NLP

The increasing effectiveness of adversarial attacks makes robust countermeasures necessary for NLP security. Defenses against adversarial attacks in NLP are either based on (reactive) detection or (proactive) model enhancement methods. Detection and filtering methods may have limited power against sophisticated and dynamic adversarial attacks. Therefore, model enhancement methods such as AdvT, functional improvement and certification could be preferred. Among these, AdvT is based on proactive inclusion of adversarial examples in training data. For instance, SmoothLLM, uses adversarial examples during training to improve robustness of LLMs with remarkable results while requiring computational resources. Phrasing is a specific

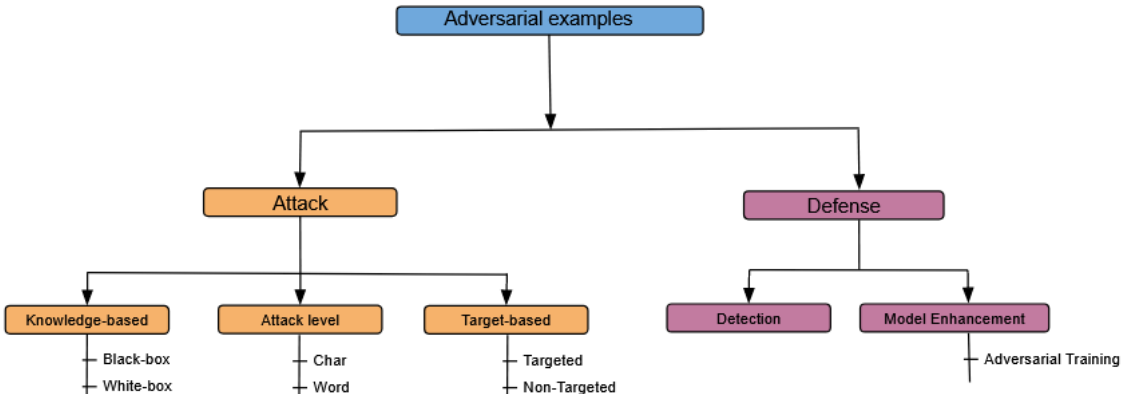

Figure 1: Taxonomy of Adversarial Attacks and Defenses in NLP.

tailored method that trains models to recognize resilient phrases. Zero-Shot Defender for Adversarial Sample Detection and Restoration (ZDDR) combines AdvT with zero-shot learning to detect and restore adversarial inputs. Adv-HotFlip, FreeLB, FreeLB++ enhance model robustness against specific attacks like word substitution at the cost of additional fine-tuning. Text purification (BERT, RoBERTa) refines representations using adversarial examples.

Input preprocessing and data transformation methods include "Synonym Encoding" which replaces words with synonyms to reduce sensitivity to specific words. Duplicate text filtering removes duplicate text to improve generalization. While it is effective, it may discard useful data. Data sanitization is based on removing sensitive information from data. It protects privacy at the potential cost of altering semantics. Finally, knowledge expansion methods augment training data with external knowledge sources. They enhance understanding but their performance depends on the quality and relevance of the knowledge added. While these defense techniques help improve NLP model robustness, increased computational complexity and vulnerabilities to specific attacks are among current limitations. Table 2 presents an overview of NLP defenses against adversarial attacks ranging from char to sentence levels.

3.3. Practical Guidance

Practitioners can create and craft adversarial attacks to assess the robustness of their model. In order to create gradient-based adversarial attacks, following steps can be utilized:

1. **Model Selection**: Choose the target model
2. **Gradient Calculation and Perturbation Generation**: Calculate gradients to generate perturbations that fool the model. Tools like `nlpaug` can help specify the attack level.
3. **Evaluation**: Test the adversarial example by comparing the model's output with the original.

An alternative option is to utilize genetic algorithms that evolve adversarial examples through natural selection. Their standard steps include initialization (generation of initial adversarial examples), evaluation (assessing their effectiveness), selection (choosing high performing examples) and crossover and mutation (creating new examples). Libraries like `NLPAug`, `TextAttack`, `Foolbox`, and `CleverHans` provide pre-implemented algorithms for crafting adversarial examples.

Transformer models (e.g., BERT, GPT-3) are generally more robust but especially their smaller variants are still susceptible to adversarial attacks. Attacks could happen in the form of word substitution (synonyms) as displayed in Figure 2.

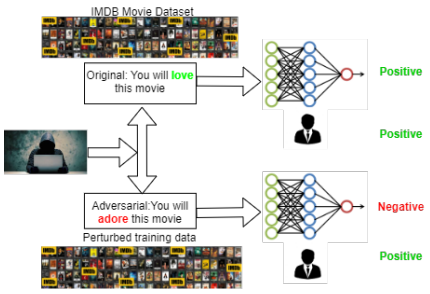

Figure 2: Adversarial example in sentiment analysis on IMDB dataset (Shaw et al. (2024))

Guo et al. (2021) presents an overview of adversarial attack techniques on deep learning models, while Hartl et al. (2020) and Y. Cheng et al. (2019) focus on RNNs and transformers based NMTs, respectively. In order to provide practical insights related to the impact of adversarial attacks on NLP methods, we provide some computational results using well known benchmark data sets of IMDB and Twitter via Table 3. The varying impacts of attacks on

Table 1: Review Highlights of Adversarial Attacks in NLP.

Paper/Year	Knowledge	Target	Level	Method under Attack	Attack Type	Data	Results
Gao et al. (2018)	Black-box	Non Targeted	Char	DeepWordBug	Word-LSTM, Char-CNN	AG News, Amazon Review Full and Polarity, DBPedia, Yahoo Answers, Yelp Review Full and Polarity, Enron Spam Email	Efficient adversarial modifications on the input tokens without gradient guidance
Gil et al. (2019)	Black-box	Non Targeted	Char	NN	HOTFLIP, DistFlip	Toxic Comment	Employ white-to-black distillation techniques to enhance adversarial attack efficiency.
Glockner et al. (2018)	Black-box	Non Targeted	Word	pre-trained GloVe embeddings	Lexically challenging sentences	SNLI	Generation of new LNI data simpler than SNLI with limited generalization
Behjati et al. (2019)	White-box	Targeted	Word	LSTM	Gradient projection based universal perturbations	AGNews, Stanford Sentiment	Effective data-independent attacks
Y. Cheng et al. (2019)	White-box	Targeted	Word	NMT	Gradient-based transformer AdvGen	LDC, NIST 2006, WMT14	Generated adversarial examples later to enhance NMT robustness with doubly adversarial inputs.
H. Zhang et al. (2020)	Binary	Targeted	Word	DNN	Metropolis-Hastings with gradient guided proposal	IMDB, SNLI	Efficient adversarial attacks
Jin et al. (2020)	Black-box	Non Targeted	Word	CNN, LSTM, BERT	TEXTFOOLER	AGNews, IMDB, Fake Yelp, MR, SNLI, MultiNLI	Effective model attacks that maintains semantic content
M. Cheng et al. (2020)	White-box	Binary	Word	seq2seq NN	Seq2Sick (gradient descent with novel loss functions)	DUC2003, DUC2004, Gigaword, WMT15	Effective attacks
Yang et al. (2020)	Black-box	Non Targeted	Word	CNN	(Probabilistic) greedy and gumbel attacks	IMDB, Yahoo! Answers	Greedy attacks are effective and and gumbel attacks are efficient on (discrete) text classifiers.
Zou et al. (2019)	White-box	Non Targeted	Word	RNN-search and transformer based NMT	Reinforcement learning with a discriminator	WMT14, LDC	Generated stable adversarial examples while maintaining semantic integrity
Zou et al. (2024)	Binary	Targeted	Word	Retrieval Augmented Generation (RAG)	PoisonedRAG	NQ, HotpotQA, MS-MARCO	Achieved 90% success with few poisoned texts highlighting vulnerability
Wallace et al. (2019)	White-box	Targeted	Sentence	ElasticSearch, RNN	Human-in-the-loop generation, question categorization, model evaluation.	Quizbowl	High impact of human-authored adversarial questions on QA models

accuracy and F1 scores can be recognized.

As Table 2 indicates, adversarial training has become popular in defenses. One key insight from this study is that AdvT not only enhances model robustness but also improves generalization to unseen adversarial examples, making the NLP system more reliable in real-world applications. This approach offers practitioners a practical pathway to fortify NLP models against increasingly sophisticated attacks, ensuring more secure and dependable performance.

Datasets with adversarial examples, e.g., word substitutions, character-level perturbations, are crucial for training and testing NLP models' robustness. Evaluating models on diverse datasets helps researchers understand their robustness and improve defense strategies and model architectures. Therefore, Tables 1,2 list the utilized data-sets in adversarial NLP literature. Several key metrics are used to evaluate the impact of adversarial attacks and defenses in NLP. These metrics help assess how well different attack techniques, defense mechanisms, and models work against adversarial threats. While accuracy is among the measures used to quantify the proportion of correctly classified examples, attack success rate measures the effectiveness of adversarial attacks. Robustness measures a model's ability to maintain performance when facing adversarial perturbations. Transferability assesses if adversarial examples for one model can deceive other models, indicating shared vulnerabilities. While text domain lacks universal benchmarks or data sets as introduced in image domain, there has been recent work such as Adversarial GLUE (B. Wang et al. (2021)) and MITRE ATLAS Matrix [1] to address that.

4. Emerging Areas & Future Directions

The increasing complexity and popularity of NLP applications emphasize the motivation for adversarial NLP frameworks. This section provides an overview of some emerging areas and potential future directions.

The integrity, reliability and robustness of

[1] https://atlas.mitre.org/matrices/ATLAS

Table 2: Review highlights of adversarial defenses in NLP.

Paper/Year	Knowledge	Target	Level	Method under Attack	Attack Type	Defense	Data	Findings
Belinkov and Bisk (2017)	Black-box	Non Targeted	Char	CNN	Natural and artificial noise	Structure invariant representation and AdvT	WCPC, RWSE, MERLIN, MAE	Increased model robustness other than faced with nuanced human errors
Sato et al. (2018)	Black-box	Non Targeted	Word	LSTM	Adversarial Perturbations	AdvT	IMDB, RCV1, Elec, MR, Dbpedia	Interpretable AdvT in NMT
Zang et al. (2019)	Black-box	Targeted	Word	BiLSTM, BERT	Semantic word substitution and particle swarm optimization	AdvT	IMDB, SST, SNLI	Superior adversarial examples and improved robustness
Maheshwary et al. (2021)	White-box	Targeted	Word	DNN	Population-based tailored optimization	AdvT with data augmentation	AGNews, IMDB, MR, Yelp, SNLI, MultiNLI	Improved resilience
Yoo and Qi (2021)	White-box	Non Targeted	Word	BERT, RoBERTa	Word substitution	A2T (Vanilla AdvT)	IMDB, MR, Yelp, SNLI	Word replacements by selecting top-k nearest neighbors in a counter-fitted word embedding for improved robustness
X. Wang et al. (2021)	Black-box	Non Targeted	Word	Word-CNN, LSTM, Bi-LSTM, BERT	Synonym substitution	Synonym Encoding Method with encoder insertion	IMDB, AGNews, Yahoo!News	SEM effectively blocks synonym substitution attacks.
Robey et al. (2023)	Black-box	Targeted	Word	LLM	Jailbreak attacks (GCG, PAIR, RANDOMSEARCH, AMPLEGCG	SmoothLLM (Duplicated randomly perturbed input prompts)	Behaviour Dataset	Lower attack success rate
Moraffah et al. (2024)	Black-box	Targeted	Word	BERT, RoBERTa, LLMs	TextFooler, TextAttack	LLM-based Adversarial purification methods	IMDB, AGNews	Improved classifier accuracy
L. Li and Qiu (2021)	White-box	Non Targeted	Token	BERT and ALBERT	Token-level accumulated perturbations	Token-Aware Virtual AdvT and normalization ball	AG News, IMDB, ConLL2003 NER, Ontonotes5.0 NER	Improved performance
Chen et al. (2024)	Black-box	Binary	Sentence	LLM	Combined log probability and LLM score, prompts for restoration	ZDDR	IMDB, SST2, AGNews	Improved detection and classification efficacy post-restoration
Y. Wang and Bansal (2018)	Black-box	Targeted	Sentence	BSAE (BiDAF + Self-Attn + ELMo)	AddSentDiverse	AdvT with semantic-relations knowledge	SQuAD	Improved machine comprehension with semantic relationship enhancements

Table 3: Select empirical results before and after attack against ensembles of CNN and BiLSTM

Models	Datasets	Before Attack		After Attack	
		Accuracy	F1 Score	Accuracy	F1 Score
BiLSTM	IMDB Movie	0.8	0.81	0.5	0.38
	Twitter	0.69	0.68	0.49	0.55
CNN+BiLSTM	IMDB Movie	0.83	0.82	0.5	0.51
	Twitter	0.72	0.74	0.51	0.61
BiLSTM+CNN	IMDB Movie	0.77	0.79	0.51	0.41
	Twitter	0.67	0.68	0.51	0.56
CNN+BiLSTM+CNN	IMDB Movie	0.77	0.79	0.5	0.41
	Twitter	0.72	0.74	0.5	0.6

development and deployment of responsible NLP based frameworks is a fundamental area of interest. Adversarial robustness frameworks are developed to evaluate and improve NLP model robustness. In addition to AdvT scaling attempts, emerging defense methods include functional improvement that involves enhancing the model's architecture and certification that provide formal guarantees of a model's robustness against specific types of adversarial attacks. For instance, refining word embeddings or incorporating attention mechanisms can make the model more robust by improving its ability to discern and mitigate adversarial perturbations. By employing methods such as randomized smoothing or robust optimization, these techniques ensure that the model's predictions remain stable within certain predefined bounds, offering practitioners a verifiable level of security. Incorporation of ethical guidelines focusing on transparency, fairness, and accountability are considered while reducing the impact of adversarial attacks. With respect to model interpretability, techniques such as attention visualization and saliency maps improve understanding of model behavior and identify vulnerabilities. In terms of tools, more adversarial attack detection and robustness testing frameworks are made public (Bird et al. (2000) and Wymberry and Jahankhani (2024)). These efforts aim to enhance the security, reliability, and trustworthiness of NLP systems, mitigating the risks posed by adversarial attacks. Increasing academia-industry collaboration presents opportunities in terms of research partnerships and data sharing initiatives.

Use of contextual information for generation of adversarial attacks is an emerging area of interest powered by the emergence of LLMs. For instance, semantic attacks such as synonym substitution manipulate the meaning of text inputs without changing their coherence, leading to plausible but incorrect predictions. Methods based on word

embeddings and syntax trees, as well as genetic algorithms and reinforcement learning are used to craft sophisticated adversarial examples that are harder to detect. In particular, black-box attacks have become more popular given the lack of knowledge about methods. This emphasizes the importance of transferability and generalization of attacks across models and domains. On the defense side, understanding generalization characteristics and developing tailored context-aware strategies is crucial for countering these attacks. In terms of domain adaptation, robustness against attacks in different real-world scenarios with diverse language characteristics is an emerging area.

4.1. Bayesian methods for adversarial NLP

Bayesian methods are increasingly relevant in adversarial machine learning (AML) due to their robustness and ability to quantify uncertainty (Rios Insua et al. (2023)). They can be used to generate attacks as well as detecting and defending against adversarial attacks. Bayesian sequence models can generate text while maintaining a measure of uncertainty, helping to avoid nonsensical or adversarially induced outputs. Similarly, Monte Carlo dropout could be used to approximate Bayesian inference in deep learning while providing uncertainty estimates. High uncertainty of Bayesian predictions may indicate possible adversarial manipulations (Zhao et al. (2020)). For instance, using Bayesian neural networks, where weights are treated as distributions rather than point estimates, can produce more reliable confidence estimates, making the system more robust to attacks that exploit overconfident but incorrect predictions. By continuously updating the model with new data, Bayesian approaches can also adapt to evolving threats, maintaining the security of the NLP system over time. Bayesian optimization can be used to tune hyperparameters or model architectures to find configurations that are less susceptible to adversarial attacks. In addition, Bayesian methods inherently provide regularization, which can make models more robust to perturbations. Bayesian generative models, such as variational autoencoders (Doersch (2016)), can generate adversarial text samples by sampling from the latent space, providing a range of AdvT examples.

Next, we introduce a relatively less explored utilization of Bayesian methods as a potential future direction. Bayesian decision theoretic designs could help model the interactions among the decision makers within an adversarial NLP context. For instance, adversarial risk analysis (ARA) (Banks et al. (2022)) approaches the decision problem/game from the maximizing expected utility perspective of a given player. It can take into account aleatory, epistemic, and solution-concept uncertainties through Bayesian models and subjective distributions for opponents' goals, capabilities, and strategies. It could provide decision support for the NLP defender while relaxing the common knowledge assumption. Such utilization of ARA in AML has been mostly limited to attacks against specific supervised algorithms (Naveiro et al. (2019)) and HMMs (Caballero et al. (2024)).

Figure 3 displays an example bi-agent influence diagram (Ekin et al. (2023)). Assume a Defender (NLP method) (D) chooses her defense $d \in D$, which is observed by an Attacker (A, he), who then chooses his attack $a \in A$, where D and A are respective sets of feasible alternatives. The consequences of the interaction for both agents depend on a random outcome $\theta \in \Theta$, with Θ the space of outcomes. Arc D-A reflects that the Attacker observes the Defender's decision. Both decision makers have their own assessment of the outcome probability, while their utility functions are functions of their own decisions, and the outcome. ARA facilitates the defender to acknowledge her uncertainty about the attacker's decision for a given defender's decision, which could be denoted as $p_D(a|d)$. This is retrieved by solving for the influence diagram in Figure 3b, and incorporated into the Defender's own decision problem displayed in Figure 3a.

Overall, Bayesian methods provide a natural framework for making decisions under uncertainty, which is crucial for adversarial defenses. Their inherent regularization and uncertainty estimation could help in detecting and resisting adversarial attacks. Prior knowledge can be integrated to guide the model towards safer predictions in the presence of adversarial inputs. Despite these advantages, computational cost and implementation complexity made them less popular compared to more traditional methods. Given the increasing interest in NLP models such as large language models, more widespread use of Bayesian adversarial NLP could be beneficial in certain settings.

5. Conclusion

This critical review provides an examination of adversarial attacks and defenses in NLP, highlighting the substantial challenges and proposing potential future directions. The rapidly evolving landscape of NLP, driven by sophisticated machine learning models, has simultaneously seen an increase in the complexity and efficacy of adversarial attacks. These attacks exploit vulnerabilities in NLP systems, leading to significant concerns regarding their reliability and security. This review highlights some of the recent attacks and defenses while providing practical guidance and coverage of emerging techniques such as Bayesian methods.

Our review identifies several key challenges in

(a) Defender's decision problem. (b) Defender analysis of Attacker problem

Figure 3: Influence diagrams for Defender and Attacker problems (Ekin et al. (2023))

addressing adversarial threats. The diversity of attack methods underscores the complexity of developing robust defenses. Moreover, the trade-off between model accuracy and robustness remains a critical issue, with many defensive strategies potentially degrading model performance. Another major challenge is the lack of standardized evaluation metrics and benchmarks, making it difficult to assess and compare the effectiveness of different defensive techniques comprehensively. Looking ahead, we have identified several future directions as critical for advancing the field. There is a pressing need for the development of more resilient NLP models that perform well in both ideal conditions while remaining reliable when challenged by adversarial attacks. This includes research into hybrid models that combine multiple defense strategies to cover a broader range of attack vectors. Another promising area is the integration of human-in-the-loop approaches, where human expertise is leveraged to detect and mitigate adversarial threats in real-time.

While significant progress has been made in understanding and mitigating adversarial attacks in NLP, the field remains in its nascent stages. Continued interdisciplinary research, combining insights from ML, cybersecurity, experts, and linguistics, will be essential in developing robust NLP systems capable of withstanding adversarial challenges. AI risk management and trustworthy AI frameworks may benefit from consideration of Bayesian methods. Bayesian decision making can help provide context-based risk mitigation while handling uncertainty and incorporating expert feedback in the form of a prior. While these frameworks are applicable for all NLP methods including topic models and LLMs; the trade-offs among model complexity, interpretability, and computational demands should be considered (Doshi-Velez and Kim (2017)). We are currently exploring the feasibility of using Bayesian methods for attacks and defenses for BERT models.

Acknowledgements

This material is based upon work supported by the Air Force Scientific Office of Research awards FA-9550-21-1-0239 and FA8655-21-1-7042. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*, 102131.

Alsmadi, I., Aljaafari, N., Nazzal, M., Alhamed, S., Sawalmeh, A. H., Vizcarra, C. P., Khreishah, A., Anan, M., Algosaibi, A., Al-Naeem, M. A., et al. (2022). Adversarial machine learning in text processing: A literature survey. *IEEE Access*, *10*, 17043–17077.

Banks, D., Gallego, V., Naveiro, R., & Ríos Insua, D. (2022). Adversarial risk analysis: An overview. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(1), e1530.

Behjati, M., Moosavi-Dezfooli, S.-M., Baghshah, M. S., & Frossard, P. (2019). Universal adversarial attacks on text classifiers. *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7345–7349.

Belinkov, Y., & Bisk, Y. (2017). Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., & Liberman, M. (2000). Atlas: A flexible and extensible architecture for linguistic annotation. *arXiv preprint cs/0007022*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Brocke, J. v., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *ECIS 2009 Proceedings*, 161.

Caballero, W. N., Camacho, J. M., Ekin, T., & Naveiro, R. (2024). Manipulating hidden-Markov-model inferences by corrupting batch data. *Computers & Operations Research*, *162*, 106478.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Chen, M., He, G., & Wu, J. (2024). ZDDR: A zero-shot defender for adversarial samples detection and restoration. *IEEE Access*.

Cheng, M., Yi, J., Chen, P.-Y., Zhang, H., & Hsieh, C.-J. (2020). Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI conference on artificial intelligence*, *34*(04), 3601–3608.

Cheng, Y., Jiang, L., & Macherey, W. (2019). Robust neural machine translation with doubly adversarial inputs. *arXiv preprint arXiv:1906.02443*.

Chien, J.-T. (2019). Deep bayesian natural language processing. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 25–30.

Cohen, S. (2022). *Bayesian analysis in natural language processing*. Springer Nature.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Dong, H., Dong, J., Yuan, S., & Guan, Z. (2022). Adversarial attack and defense on natural language processing in deep learning: A survey and perspective. *International conference on machine learning for cyber security*, 409–424.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ekin, T., Naveiro, R., Insua, D. R., & Torres-Barrán, A. (2023). Augmented probability simulation methods for sequential games. *European Journal of Operational Research*, *306*(1), 418–430.

Esmradi, A., Yip, D. W., & Chan, C. F. (2023). A comprehensive survey of attack techniques, implementation, and mitigation strategies in large language models. *International Conference on Ubiquitous Security*, 76–95.

Gao, J., Lanchantin, J., Soffa, M. L., & Qi, Y. (2018). Black-box generation of adversarial text sequences to evade deep learning classifiers. *2018 IEEE Security and Privacy Workshops (SPW)*, 50–56.

Gil, Y., Chai, Y., Gorodissky, O., & Berant, J. (2019). White-to-black: Efficient distillation of black-box adversarial attacks. *arXiv preprint arXiv:1904.02405*.

Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Goyal, S., Doddapaneni, S., Khapra, M. M., & Ravindran, B. (2023). A survey of adversarial defenses and robustness in NLP. *ACM Computing Surveys*, *55*(14s), 1–39.

Guo, C., Sablayrolles, A., Jégou, H., & Kiela, D. (2021). Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*.

Hartl, A., Bachl, M., Fabini, J., & Zseby, T. (2020). Explainability and adversarial robustness for RNNs. *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*, 148–156.

Huang, H., Kajiwara, T., & Arase, Y. (2021). Definition modelling for appropriate specificity. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2499–2509.

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*(05), 8018–8025.

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural language processing: History, evolution, application, and future work. *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020*, 365–375.

Li, L., & Qiu, X. (2021). Token-aware virtual adversarial training in natural language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(9), 8410–8418.

Li, X., Qiu, K., Qian, C., & Zhao, G. (2020). An adversarial machine learning method based on opcode n-grams feature in malware

detection. *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, 380–387.

Maheshwary, R., Maheshwary, S., & Pudi, V. (2021). Generating natural language attacks in a hard label black box setting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(15), 13525–13533.

Moraffah, R., Khandelwal, S., Bhattacharjee, A., & Liu, H. (2024). Adversarial text purification: A large language model approach for defense. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 65–77.

Naveiro, R., Redondo, A., Insua, D. R., & Ruggeri, F. (2019). Adversarial classification: An adversarial risk analysis approach. *International Journal of Approximate Reasoning*, *113*, 133–148.

Qiu, S., Liu, Q., Zhou, S., & Huang, W. (2022). Adversarial attack and defense technologies in natural language processing: A survey. *Neurocomputing*, *492*, 278–307.

Rios Insua, D., Naveiro, R., Gallego, V., & Poulos, J. (2023). Adversarial machine learning: Bayesian perspectives. *Journal of the American Statistical Association*, *118*(543), 2195–2206.

Robey, A., Wong, E., Hassani, H., & Pappas, G. J. (2023). Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

Sato, M., Suzuki, J., Shindo, H., & Matsumoto, Y. (2018). Interpretable adversarial perturbation in input embedding space for text. *arXiv preprint arXiv:1805.02917*.

Shaw, L., Wasim Ansari, M., & Ekin, T. (2024). Bertguard: Robust text classification against adversarial attacks. *Techrxiv preprint*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wallace, E., Rodriguez, P., Feng, S., Yamada, I., & Boyd-Graber, J. (2019). Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, *7*, 387–401.

Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A. H., & Li, B. (2021). Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Wang, R. (2022). Evaluation of four black-box adversarial attacks and some query-efficient improvement analysis. *2022 Prognostics and Health Management Conference (PHM-2022 London)*, 298–302.

Wang, X., Hao, J., Yang, Y., & He, K. (2021). Natural language adversarial defense through synonym encoding. *Uncertainty in Artificial Intelligence*, 823–833.

Wang, Y., & Bansal, M. (2018). Robust machine comprehension models via adversarial training. *arXiv preprint arXiv:1804.06473*.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii–xxiii.

Wymberry, C., & Jahankhani, H. (2024). An approach to measure the effectiveness of the mitre atlas framework in safeguarding machine learning systems against data poisoning attack. In *Cybersecurity and artificial intelligence: Transformational strategies and disruptive innovation* (pp. 81–116). Springer.

Yang, P., Chen, J., Hsieh, C.-J., Wang, J.-L., & Jordan, M. I. (2020). Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *Journal of Machine Learning Research*, *21*(43), 1–36.

Yoo, J. Y., & Qi, Y. (2021). Towards improving adversarial training of nlp models. *arXiv preprint arXiv:2109.00544*.

Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., & Sun, M. (2019). Word-level textual adversarial attacking as combinatorial optimization. *arXiv preprint arXiv:1910.12196*.

Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R., & King, I. (2018). Topic memory networks for short text classification. *arXiv preprint arXiv:1809.03664*.

Zhang, H., Zhou, H., Miao, N., & Li, L. (2020). Generating fluent adversarial examples for natural languages. *arXiv preprint arXiv:2007.06174*.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1–41.

Zhao, R., Su, H., & Ji, Q. (2020). Bayesian adversarial human motion synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6225–6234.

Zou, W., Geng, R., Wang, B., & Jia, J. (2024). Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.

Zou, W., Huang, S., Xie, J., Dai, X., & Chen, J. (2019). A reinforced generation of adversarial examples for neural machine translation. *arXiv preprint arXiv:1911.03677*.