Data & Policy (2025), 7: e64
doi:10.1017/dap.2025.10030

RESEARCH ARTICLE

Adversarial natural language processing: overview, challenges, and policy implications

Laxmi Shaw, Mohammed Wasim Ansari and Tahir Ekin

Department of Information Systems and Analytics, Texas State University, San Marcos, TX, USA
Corresponding author: Tahir Ekin; Email: t_e18@txstate.edu

Received: 11 April 2025; **Revised:** 30 June 2025; **Accepted:** 14 August 2025

Keywords: adversarial machine learning; AI security; Bayesian methods; natural language processing; text classification

Abstract

The emergence of large language models has significantly expanded the use of natural language processing (NLP), even as it has heightened exposure to adversarial threats. We present an overview of adversarial NLP with an emphasis on challenges, policy implications, emerging areas, and future directions. First, we review attack methods and evaluate the vulnerabilities of popular NLP models. Then, we review defense strategies that include adversarial training. We describe major policy implications, identify key trends, and suggest future directions, such as the use of Bayesian methods to improve the security and robustness of NLP systems.

Policy Significance Statement

This work highlights the impact of adversarial attacks on natural language processing (NLP) systems, especially in high-stakes application domains such as healthcare. As these artificial intelligence (AI) methods become more powerful, policymakers must ensure that they are used fairly, securely, and transparently. Key concerns include preventing bias, protecting privacy, and managing the high-energy demands of large-scale models. This paper explores attacks, defenses, and the growing role of Bayesian methods to improve robustness and decision-making. However, these advances also raise concerns about data protection and algorithmic bias. Policymakers should promote transparency, ethical standards, and sustainable AI practices. Balanced regulation will allow NLP technologies to remain trustworthy, effective, and aligned with public interest across sectors and international boundaries.

1. Introduction

Natural language processing (NLP) has undergone significant evolution over the years. 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 revolution 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 bidirectional encoder representations from transformers (BERT) and generative pre-trained transformers

This research article was awarded Open Data for transparent practices. See the Data Availability Statement for details.

(GPT) have yielded cutting-edge performance in tasks such as language understanding, generation, translation, and summarization (Johri et al., 2021). NLP-based technologies enhance various fields such as healthcare, customer service, education, and entertainment (Esmradi et al., 2023). This rapid progress and increasing number of NLP applications in human-facing domains emphasize the importance of addressing relevant policy implications to ensure an equitable, ethical, and effective deployment.

Moreover, increasing use of NLP systems amplifies the security concerns due to potential data manipulation that can 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). Integrating adversarial testing and evaluation into the development lifecycle helps uncover and address vulnerabilities before deploying NLP systems in real-world scenarios. However, securing these systems remains a persistent challenge, demanding ongoing innovation and adaptability. Emerging vulnerabilities, especially in high-stakes domains like healthcare, law, and cybersecurity, raise complex policy issues. To mitigate risks from misleading or harmful outputs, a strong regulatory framework is essential. This includes safeguards such as secure model training, continuous adversarial testing, and transparency throughout model deployment.

Several review papers have surveyed adversarial attacks and defenses in NLP focusing on different aspects. Zhang et al. (2020b) provided 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 emphasized the diversity of attack methods and their impact on various NLP tasks. Li et al. (2020) focused 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) covered both adversarial attacks and defenses on NLP in deep learning. Alsmadi et al. (2022) presented a survey of methods for text generation. Cheng et al. (2019) examined 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 defense mechanisms are not fully captured in these surveys, particularly empirical comparisons involving ensemble techniques developed in the last few years. They often lack guidance on the practical implementations, making it challenging for practitioners to apply these methods. In terms of LLM defenses, Esmradi et al. (2023) analyzed LLM security vulnerabilities and reviewed effective defense strategies including data sanitization, encryption-based methods, differential privacy, and filtering. The review of Qiu et al. (2022) covered 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. Despite broad coverage in recent literature; practical guidance, emerging techniques, such as Bayesian methods and overarching policy implications, remain insufficiently explored in this fast-moving field of study.

We address these limitations by extending Shaw et al. (2025) with an overview of policy implications related to adversarial NLP. In particular, we review 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, in addition to policy implications. 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 policy implications and 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 and existing defenses, while providing practical guidance. Section 4 discusses the policy implications and the

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

relevance of adversarial NLP. Section 5 presents emerging areas and future research directions. The manuscript concludes with final remarks in Section 6.

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.

The emerging use of NLP emphasizes challenges and policy implications related to data privacy, fairness, bias mitigation, and ethical AI governance. First, the reliance of transformer-based models like BERT and GPT on large-scale datasets raises concerns regarding user data privacy, especially when training on sensitive information (Carlini et al., 2021). Regulatory frameworks such as the General Data Protection Regulation (GDPR) emphasize the need for data anonymization, consent-based collection, and user control over personal information, which may conflict with the vast, unregulated data sources often used in pretraining these models. Furthermore, bias in training data can propagate into model outputs, leading to unfair or discriminatory outcomes in applications such as automated hiring, content moderation, and sentiment analysis (Bender et al., 2021). Additionally, the increasing carbon footprint of large-scale NLP models due to the computational needs raises sustainability concerns. This encourages the use of training techniques such as pruning, quantization, and knowledge distillation that could reduce energy consumption (Strubell et al., 2020). Addressing these challenges requires a balanced regulatory approach that promotes innovation while ensuring ethical AI deployment in real-world applications. The use of NLP in adversarial environments amplifies some of these challenges, resulting in elevated policy implications (Schlarmann and Hein, 2023).

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

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

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 complexity and computational cost may limit the real-world adoption. Bayesian methods in NLP could also present policy challenges related to fairness and transparency. Bayesian inference inherently depends on prior probabilities, which can introduce bias if the training data are not representative (Evans and Guo, 2021). This is particularly relevant in adversarial NLP, where robustness of the models against misinformation, spam, and adversarial attacks, is crucial.

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 centrally 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 Xplore," 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 to weaken the performance and reliability of NLP systems. These methods usually involve changing text data, at

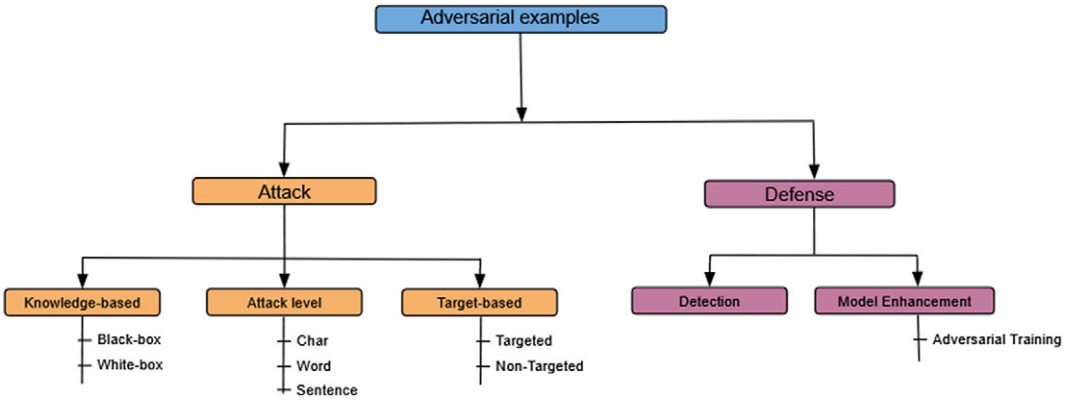

Figure 1. *Taxonomy of adversarial attacks and defenses in NLP.*

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

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 correspond 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 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 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 (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). Guo et al. (2021) presents an overview of adversarial attack techniques on deep learning models, while Hartl et al. (2020) and Cheng et al. (2019) focus on RNNs and transformers based NMTs, respectively. Smaller variants of the transformer models (e.g., BERT, GPT-3) are more susceptible to adversarial attacks. Table 1 presents an overview of highlights of adversarial attacks, ranging from char-level to sentence-level attacks, against various NLP models.

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 the robustness of LLMs with remarkable results while requiring computational resources. Phrasing is a specific 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. As shown in Table 2, several other methods enhance model robustness against specific attacks at the cost of additional fine-tuning. 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. Despite its effectiveness, it may also 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.

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

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	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
Cheng et al. (2019)	White-box	Targeted	Word	NMT	Gradient-based transformer AdvGen	LDC, NIST 2006, WMT14	Adversarial examples to enhance NMT robustness with doubly adversarial inputs.
Zhang et al. (2020a)	Binary	Targeted	Word	DNN	Metropolis-Hastings with gradient guided proposal	IMDB, SNLI	Efficient adversarial attacks

(Continued)

Table 1. *Continued*

Paper/Year	Knowledge	Target	Level	Method under attack	Attack type	Data	Results
Jin et al. (2020)	Black-box	Non Targeted	Word	CNN, LSTM, BERT	TEXTFOOLER	AGNews, IMDB, Fake Yelp, MR, SNLI, MultiNLI	Effective model attacks that maintain semantic content
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	Effective greedy attacks and efficient gumbel attacks 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	Stable adversarial examples that maintain semantic integrity
Zou et al. (2024)	Binary	Targeted	Word	Retrieval Augmented Generation (RAG)	PoisonedRAG	NQ, HotpotQA, MS-MARCO	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	Impactful human-authored adversarial questions on QA models

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
Wang et al. (2021b)	Black-box	Non Targeted	Word	Word-CNN, LSTM, Bi-LSTM, BERT	Synonym substitution	Synonym Encoding Method with encoder insertion	IMDB, AGNews, Yahoo!News	Effective blocking of synonym substitution attacks.

(Continued)

Table 2. *Continued*

Paper/Year	Knowledge	Target	Level	Method under attack	Attack type	Defense	Data	Findings
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
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
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

3.3. Practical guidance

Practitioners can use various techniques to design adversarial attacks that evaluate and stress-test model robustness. These include open source libraries like `nlpaug`[1], `TextAttack`[2], `Foolbox`[3], and `CleverHans`[4] that provide algorithms for crafting adversarial examples. One widely used option is to utilize genetic algorithmic frameworks 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).

In this study, we focused on gradient-based adversarial attacks by utilizing the grand framework of model selection, gradient calculation and perturbation generation, and evaluation, through the use of open-source Python library `Nlpaug`. This data augmentation library offers methods such as synonym replacement, contextual word substitution, and back translation, making it versatile for different NLP tasks, such as sentiment analysis and topic classification. Augmenting text data has been shown to enhance the performance of NLP models, particularly for classification tasks (Bayer et al., 2023). In particular, we generated adversarial attacks on text data by applying random attacks, e.g., adding spelling errors, word splitting, and random perturbations[5]. The relevant code can be accessed at GitHub[6]. Figure 2 displays a practical example for a word substitution (synonym) attack.

Table 3 presents the impact of such adversarial attacks on different combinations of NLP methods against both IMDB and Twitter[7] datasets. The varying impacts of attacks on accuracy and F1 scores can be recognized. For instance, accuracy dropped by 30% under attack, illustrating the practical risk to deployed NLP systems.

In terms of defenses, Table 2 indicates the popularity of adversarial training (AdvT). In our practical implementations (Shaw and Ekin, 2024), we also have recognized that AdvT not only enhances the robustness of the model but also improves generalization to unseen adversarial examples. This could offer

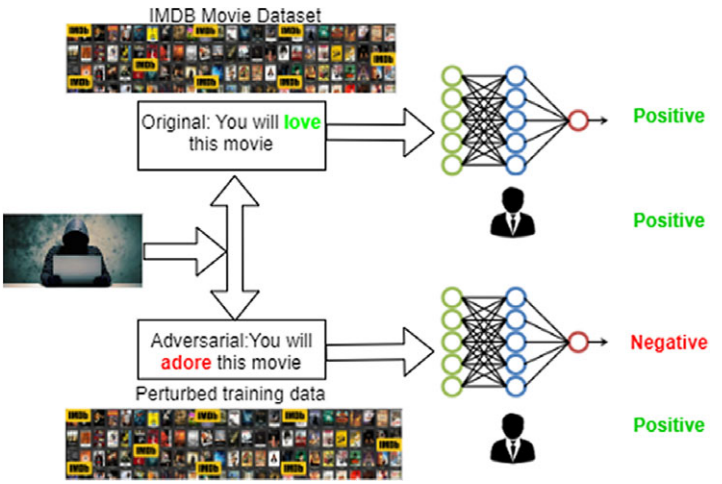

Figure 2. *Adversarial sentiment analysis example on IMDB dataset (Shaw et al., 2024).*

[1] https://github.com/makcedward/nlpaug.
[2] https://github.com/QData/TextAttack.
[3] https://github.com/bethgelab/foolbox.
[4] https://github.com/cleverhans-lab/cleverhans.
[5] https://nlpaug.readthedocs.io/en/latest/augmenter/word/word.html.
[6] https://github.com/makcedward/nlpaug/blob/master/docs/overview/overview.rst.
[7] https://github.com/Sweety176/Sentiment_datasets.

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

Table 3. *Select empirical results before and after attacks against ensembles of CNN and BiLSTM*

		Before attack		After attack	
Models	Datasets	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

practitioners a practical pathway to fortify NLP models against increasingly sophisticated attacks, ensuring more secure and reliable 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 and 2 list the utilized datasets 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 (Wang et al., 2021a) and MITRE ATLAS Matrix[8] to address that.

4. Policy implications

NLP has revolutionized the way we interact with technology, offering transformative applications in domains such as healthcare (Wong et al., 2018; Jerfy et al., 2024; Khattak and Rabbi, 2023), legal systems (Hovy and Spruit, 2016) with profound societal and economic implications. In healthcare, NLP powers applications like clinical text analysis, electronic health record processing, and patient sentiment analysis. Inaccuracies or biases in these systems can lead to incorrect diagnoses or compromised patient care (Schopow et al., 2023). Policies should ensure rigorous validation of NLP models used in healthcare and establish guidelines for their deployment. For instance, requiring official approvals for NLP-powered medical devices can help maintain high standards of safety and efficacy. NLP tools are also increasingly employed in legal document analysis, contract review, and case law research. While these applications enhance efficiency, they also raise concerns about accountability and interpretability (Aletras et al., 2016; Doshi-Velez and Kim, 2017; Ariai and Demartini, 2024). Governments may establish guidelines to ensure that NLP systems in legal contexts remain interpretable and unbiased. Collaboration with legal professionals can help create standards for ethical usage (Quevedo et al., 2023).

Adversarial NLP attacks, such as phishing or data poisoning, enhance cybersecurity risks of these systems (Story et al., 2019). These attacks could include but are not limited to data extraction, bias manipulation, and monetization of misinformation. For instance, subtle adversarial prompts targeting

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

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

LLMs integrated into customer service systems could trick models into revealing masked personal data (e.g., partial identification, account info) through carefully crafted follow-up queries. In educational or medical domains, adversarial inputs could distort model outputs to favor certain commercial products, institutions, or treatments, nudging user decisions toward sponsored or malicious ends. Monetization of misinformation could occur by manipulating an NLP assistant that could redirect users to unofficial monetized third party services. Lastly, adversaries may craft synthetic personas or requests to overwhelm public service chatbots, to prevent vulnerable users from getting timely help.

These examples emphasize how seemingly minor adversarial manipulations can lead to tangible societal risks, and could reinforce the necessity of proactive defense mechanisms. The resulting policy implications should be considered for equitable, ethical, and effective deployment. Policies requiring regular updates to NLP systems to address emerging threats and encourage the development of secure architectures are possibly needed. International coordination efforts on cybersecurity policies, such as information sharing agreements, are critical to improve global resilience.

4.1. Ethical challenges and governance

The applications of NLP systems may often result in ethical challenges related to fairness and privacy. One of the most pressing challenges in NLP is the presence of bias in training data. Models trained on biased data can inadvertently replicate and amplify discriminatory patterns. For instance, sentiment analysis systems may misclassify language from minority dialects, and automated hiring tools can exhibit gender or racial biases. Governments and organizations should possibly create standards for bias detection and mitigation. Policies could advocate for transparency in dataset composition and require regular audits of NLP models to evaluate fairness. Robust governance frameworks can help ensure that NLP technologies do not perpetuate or exacerbate societal inequities. Frameworks such as the European Union's AI Act, which emphasizes accountability and fairness, serve as important precedents for global adoption.

Moreover, privacy is a critical consideration in NLP applications, especially in domains like healthcare or legal services, where sensitive information is processed. LLMs trained on vast amounts of data may inadvertently memorize and reproduce private or confidential information. Hence, policy interventions may be essential to safeguard user data and enforce compliance with privacy regulations like the GDPR. Adopting differential privacy techniques during model training and ensuring end-to-end encryption for NLP-powered systems can help address these concerns.

These ethical challenges become even more crucial in adversarial environments, and policy makers need to be aware of the implications. Accountability mechanisms that are crucial to prevent the misuse of NLP technologies should consider potential manipulation of data and models. Organizations can increase the transparency and document their decision-making processes of their models and provide avenues for recourse in cases of harm caused by algorithmic decisions.

4.2. Policy and regulatory needs for adversarial robustness

The threat posed by adversarial attacks to the reliability and security of the NLP systems creates the potential need for regulatory frameworks. Such policies that encourage secure model development are especially crucial for high-stakes applications. We can address some of these challenges by embedding safeguard regulations as part of the development lifecycle for NLP systems. Similarly to the required cybersecurity certifications (Alawida et al., 2023), NLP models could undergo adversarial stress tests to confirm their resilience against attacks on data and models (Singh et al., 2022). Governments and public agencies can incentivize the development of more robust NLP technologies by funding research and offering tax incentives to companies that prioritize security. Public-private partnerships can also facilitate research into more effective techniques to secure NLP systems. For instance, frameworks could mandate transparency in how models handle adversarial examples, ensuring public trust and accountability (Al-Maliki et al., 2024).

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

Adversarial threats transcend borders, particularly when NLP models are deployed in multilingual or cross-cultural contexts (e.g., global customer service chatbots or translation systems). Such a global nature of NLP development and deployment may require international collaboration on policy frameworks and standardization. Differences in data privacy laws, ethical standards, and regulatory requirements between countries can hinder progress and create compliance challenges for organizations. Efforts to harmonize data privacy regulations, such as aligning GDPR with U.S. frameworks like the California Consumer Privacy Act, can streamline compliance for NLP applications operating across jurisdictions. In addition, policies promoting open data standards and interoperability can facilitate cross-border collaboration in NLP research. For example, shared datasets for adversarial training and multilingual NLP can help advance the field while adhering to ethical standards. As nations become more protective of their digital resources, policies should balance the need for digital sovereignty with the benefits of international collaboration. Agreements on data sharing and joint research initiatives can ensure mutual benefits while respecting national interests.

As LLMs integrate Bayesian techniques for probabilistic reasoning, regulatory guidelines should ensure that these models maintain ethical standards, transparency, and responsible AI governance. Encouraging open-source Bayesian NLP frameworks and federated learning approaches could further enhance privacy-preserving AI applications while maintaining model robustness and scalability. Future policies should focus on balancing innovation and regulation, ensuring that Bayesian NLP methods contribute to ethical, fair, and secure AI systems.

5. Emerging areas and future directions

The increasing popularity of NLP applications and the extent of adversarial attacks emphasize the motivation for adversarial NLP frameworks with increasing policy implications. This section provides an overview of the several emerging areas and potential future directions.

The integrity, reliability, and robustness of the development and deployment of responsible NLP-based frameworks are fundamental areas 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 provides 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 is 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; 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

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

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.

5.1. 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. Lastly, Bayesian decision-theoretic models such as adversarial risk analysis (Banks et al., 2022) could be used for AML (Ekin et al., 2023; Caballero et al., 2024). These could be beneficial to model the interactions among the decision makers within adversarial NLP contexts (Shaw, 2025).

Overall, Bayesian methods offer a principled framework for decision-making under uncertainty, an essential aspect of developing robust adversarial defenses. Their ability to incorporate prior knowledge, provide calibrated uncertainty estimates, and apply natural regularization makes them well-suited for identifying and mitigating adversarial inputs. These strengths can guide models toward safer predictions when faced with perturbed or ambiguous data. However, the high computational demands and the implementation complexity have limited their adoption relative to more conventional approaches. As interest in large-scale NLP systems, particularly LLMs, continues to grow, there is a renewed need to reassess these trade-offs and explore the broader application of Bayesian approaches in adversarial NLP settings.

5.2. Emerging policy challenges

The emergence of large-scale NLP models, particularly LLMs, has introduced significant new policy challenges and amplified existing concerns. These challenges span domains, such as national security, public trust, misinformation, access, sustainability, and regulatory oversight.

First, adversarial attacks can significantly undermine the reliability of government-run IT systems that use NLP models. For instance, LLMs deployed in public service portals, e.g., immigration, social benefits, and tax filing systems, could be manipulated into giving incorrect or misleading guidance. An adversarial input might alter eligibility decisions, misroute applications, or subtly redirect users to malicious third-party websites. In addition to affecting operational efficiency, these may also result in privacy breaches, denial of services, or legal liability for government agencies. These risks highlight the urgent need for adversarial robustness in publicly deployed AI systems.

Second, many autonomous systems (e.g., self-driving cars, automated drones) increasingly rely on NLP for tasks like interpreting commands, processing sensor data, or interacting with humans. Adversarial threats could disrupt these systems, leading to potential safety risks. This eventually could diminish public trust in NLP. Potential governmental intervention can help establish minimum robustness and transparency standards, incentivize industry practices like "adversarial audits" and robustness certification, and coordinate international policy to prevent cross-border adversarial attacks.

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

Third, training and deploying large NLP models consume substantial energy resources, contributing to environmental concerns. Policies can encourage the development of energy-efficient model architectures, sustainable deployment and environmental reporting standards in alignment with global sustainability goals.

Finally, the generative capabilities of large-scale models result in powerful tools for misinformation and propaganda. Governments may support transparency when it comes to detecting and mitigating the spread of (false) information generated by NLP systems. Public awareness campaigns can complement these efforts. This also has consequences on trust in NLP systems. As NLP systems become increasingly sophisticated, their accessibility to under-resourced communities and languages remains a concern. Policies can support initiatives to develop NLP tools for low-resource languages and to enhance equitable access to these technologies.

These implications emphasize the importance of detection mechanisms and their use for public benefit while keeping the systems accessible. Policymakers face the complex task of balancing technical defenses and accessibility with legislative oversight. Ensuring secure model development is essential, especially when models are used in sensitive domains like legal aid, medical triage, or education. Prevention and early detection require investment in both tooling and organizational workflows, including training for civil servants to identify anomalies. Legislation might also be necessary, especially in creating (international) standards for responsible deployment, redressal mechanisms for affected users, and penalties for malicious adversarial input crafting. International collaboration and proactive regulation may also help navigate the complexities of NLP systems. Establishing a globally shared knowledge hub that gathers and publishes real-time adversarial threat intelligence for NLP systems can be beneficial. This could help with standardization of adversarial robustness and help responding to adversarial incidents or failures (e.g., LLM misuse, model degradation under attack), in the spirit of cybersecurity breach notification laws. Ultimately, an integrated approach managing technical, procedural, and legal aspects is likely to be most effective for NLP governance.

6. Conclusion

This critical review provides an examination of adversarial attacks and defenses in NLP, describing the challenges and policy implications while describing 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 attacks and defenses while providing practical guidance and coverage of emerging techniques such as Bayesian methods.

Our review identifies several key challenges in 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.

The increasing reliance on NLP systems introduces critical policy challenges that need to be addressed for equitable and secure deployment. Adversarial attacks highlight the need for robust regulatory frameworks. Policymakers may focus on enforcing standards for adversarial testing to improve resilience against such attacks before deployment. International coordination can help create cross-border agreements addressing the global nature of adversarial threats, promoting a collective defense. Furthermore,

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

initiatives to incentivize private organizations and research institutions to develop secure NLP systems, such as through grants or tax benefits, can accelerate innovation in this space. These measures are pivotal for maintaining trust in NLP technologies, especially in high-stakes sensitive domains like healthcare, legal systems, and cybersecurity. As NLP continues to reshape industries and societies, a strong policy foundation will be essential to maximize its benefits while mitigating its risks.

While major progress has been made in understanding and mitigating adversarial attacks in NLP, the field remains in its nascent stages both in terms of methodological developments and policy guidelines. Continued interdisciplinary research, combining insights from ML, cybersecurity, policy 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 offers a powerful approach for context-aware risk mitigation by managing uncertainty and integrating expert knowledge through priors. These frameworks apply broadly across NLP methods, from topic models to large language models (LLMs). However, balancing model complexity, interpretability, computational efficiency, and policy implications remains an active and important area of research.

Abbreviations

AI	Artificial intelligence
BERT	Bidirectional encoder representations from transformers
CNN	Convolutional neural network
GenAI	Generative AI
GPT	Generative pre-trained transformer
LLMs	Large language models
NLP	Natural language processing
RNN	Recurrent neural network

Data availability statement. The data that support the findings of this study are open-source and made available Shaw (2025).

Author contribution. Conceptualization: LS and TE. Methodology: LS and TE. Formal analysis and investigation: LS, MWA and TE. Writing—original draft: LS and TE, Writing—review and editing: TE. Funding acquisition: TE Resources: TE Supervision: LS and TE. All authors approved the final submitted draft.

Funding statement. This work was supported by the National Science Foundation under research grant 2,334,268, the Air Force Office of Scientific Research awards FA-9550-21-1-0239 and FA8655-21-1-7042. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests. The authors declare none.

References

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

Alawida M, **Mejri S**, **Mehmood A**, **Chikhaoui B and Isaac Abiodun O** (2023) A comprehensive study of chatgpt: Advancements, limitations, and ethical considerations in natural language processing and cybersecurity. *Information 14*(8), 462.

Aletras N, **Tsarapatsanis D**, **Preoţiuc-Pietro D and Lampos V** (2016) Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science 2*, e93.

Al-Maliki S, **Qayyum A**, **Ali H**, **Abdallah M**, **Qadir J**, **Hoang DT**, **Niyato D and Al-Fuqaha A** (2024) Adversarial machine learning for social good: Reframing the adversary as an ally. *IEEE Transactions on Artificial Intelligence 5*(9), 4322–4343.

Alsmadi I, **Aljaafari N**, **Nazzal M**, **Alhamed S**, **Sawalmeh AH**, **Vizcarra CP**, **Khreishah A**, **Anan M**, **Algosaibi A**, **al-Naeem MA**, **Aldalbahi A and al-Humam A** (2022) Adversarial machine learning in text processing: A literature survey. *IEEE Access 10*, 17043–17077.

Ariai F and Demartini G (2024) Natural language processing for the legal domain: A survey of tasks,datasets, models, and challenges. arXiv preprint 2410.21306. Available at https://arxiv.org/abs/2410.21306.

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

Bayer M, **Kaufhold M-A**, **Buchhold B**, **Keller M**, **Dallmeyer J and Reuter C** (2023) Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *International Journal of Machine Learning and Cybernetics 14*(1), 135–150.

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

Behjati M, **Moosavi-Dezfooli S-M**, **Baghshah MS and Frossard P** (2019) Universal adversarial attacks on text classifiers. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton: IEEE, pp. 7345–7349. https://doi.org/10.1109/ICASSP.2019.8682430.

Belinkov Y and Bisk Y (2017) Synthetic and natural noise both break neural machine translation. arXiv preprint 1711.02173. Available at https://arxiv.org/abs/1711.02173.

Bender EM, **Gebru T**, **McMillan-Major A and Shmitchell S** (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery (ACM), pp. 610–623.

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

Blei DM, **Ng AY and Jordan MI** (2003) Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Brocke Jv, **Simons A**, **Niehaves B**, **Riemer K**, **Plattfaut R and Cleven A** (2009) Reconstructing the giant: On the importance of rigour in documenting the literature search process. In *ECIS 2009 Proceedings*. AIS Electronic Library, p. 161. Available at https://aisel.aisnet.org/ecis2009/16.

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

Carlini N, **Tramer F**, **Wallace E**, **Jagielski M**, **Herbert-Voss A**, **Lee K**, **Roberts A**, **Brown T**, **Song D**, **Erlingsson U**, **Oprea A and Raffel C** (2021). Extracting training data from large language models. In *30th Usenix Security Symposium (Usenix Security 21)*. Berkeley, CA: USENIX Association, pp. 2633–2650.

Chao P, **Robey A**, **Dobriban E**, **Hassani H**, **Pappas GJ and Wong E** (2023) Jailbreaking black box large language models in twenty queries. arXiv preprint 2310.08419. Available at https://arxiv.org/abs/2310.08419.

Chen M, **He G and Wu J** (2024) *ZDDR: A Zero-Shot Defender for Adversarial Samples Detection and Restoration*. Piscataway, NJ: IEEE Access.

Cheng Y, **Jiang L and Macherey W** (2019) Robust neural machine translation with doubly adversarial inputs. arXiv preprint 1906.02443. Available at https://arxiv.org/abs/1906.02443.

Cheng M, **Yi J**, **Chen P-Y**, **Zhang H and Hsieh C-J** (2020) Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. *34*. Washington, DC: Association for Advancement of Artificial Intelligence, pp. 3601–3608.

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

Cohen S (2022) *Bayesian Analysis in Natural Language Processing*. Cham: Springer Nature.

Devlin J, **Chang M-W**, **Lee K and Toutanova K** (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 1810.04805. Available at https://arxiv.org/abs/1810.04805.

Doersch C (2016) Tutorial on variational autoencoders. arXiv preprint 1606.05908. Available at https://arxiv.org/abs/1606.05908.

Dong H, **Dong J**, **Yuan S and Guan Z** (2022) Adversarial attack and defense on natural language processing in deep learning: A survey and perspective. In *International Conference on Machine Learning for Cyber Security*. Heidelberg, Germany: Springer Nature, pp. 409–424.

Doshi-Velez F and Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint 1702.0868. Available at https://arxiv.org/abs/1702.08608.

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

Esmradi A, **Yip DW and Chan CF** (2023) A comprehensive survey of attack techniques, implementation, and mitigation strategies in large language models. In *International Conference on Ubiquitous Security*. Heidelberg, Germany: Springer Nature, pp. 76–95.

Evans M and Guo Y (2021) Measuring and controlling bias for some Bayesian inferences and the relation to frequentist criteria. *Entropy 23*(2), 190.

Gao J, **Lanchantin J**, **Soffa ML and Qi Y** (2018) Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*. New York City, NY: IEEE, pp. 50–56.

Gil Y, **Chai Y**, **Gorodissky O and Berant J** (2019) White-to-black: Efficient distillation of black-box adversarial attacks. arXiv preprint 1904.02405. Available at https://arxiv.org/abs/1904.02405.

Glockner M, **Shwartz V and Goldberg Y** (2018) Breaking NLI systems with sentences that require simple lexical inferences. arXiv preprint 1805.02266. Available at https://arxiv.org/abs/1805.02266.

Goyal S, **Doddapaneni S**, **Khapra MM and 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 and Kiela D** (2021) Gradient-based adversarial attacks against text transformers. arXiv preprint 2104.13733. Available at https://arxiv.org/abs/2104.13733.

Hartl A, **Bachl M**, **Fabini J and Zseby T** (2020) Explainability and adversarial robustness for RNNs. In *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (Bigdataservice)*. New York City, NY: IEEE, pp. 148–156.

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

Hovy D and Spruit SL (2016) The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 591–598.

Huang H, **Kajiwara T and Arase Y** (2021) Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2499–2509.

Jerfy A, **Selden O and Balkrishnan R** (2024) The growing impact of natural language processing in healthcare and public health. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing 61*, 00469580241290095.

Jin D, **Jin Z**, **Zhou JT and Szolovits P** (2020) Is bert really robust? A strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. *34*. Washington, DC: Association for the Advancement of Artificial Intelligence, pp. 8018–8025.

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

Khattak WA and Rabbi F (2023) Ethical considerations and challenges in the deployment of natural language processing systems in healthcare. *International Journal of Applied Health Care Analytics 8*(5), 17–36.

Li L and Qiu X (2021) Token-aware virtual adversarial training in natural language understanding. In *Proceedings of the Aaai Conference on Artificial Intelligence*, Vol. *35*. Washington, DC: Association for Advancement of Artificial Intelligence, pp. 8410–8418.

Li X, **Qiu K**, **Qian C and Zhao G** (2020) An adversarial machine learning method based on opcode n-grams feature in malware detection. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*. New York City, NY: IEEE, pp. 380–387.

Maheshwary R, **Maheshwary S and Pudi V** (2021) Generating natural language attacks in a hard label black box setting. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. *35*. Washington, DC: Association for Advancement of Artificial Intelligence, pp. 13525–13533.

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

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

Quevedo E, **Cerny T**, **Rodriguez A**, **Rivas P**, **Yero J**, **Sooksatra K and Taibi D** (2023) Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications. *IEEE Access 12*, 145286–145317.

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

Robey A, **Wong E**, **Hassani H and Pappas GJ** (2023) Smoothllm: Defending large language models against jailbreaking attacks. arXiv preprint 2310.03684. Available at https://arxiv.org/abs/2310.03684.

Sato M, **Suzuki J**, **Shindo H and Matsumoto Y** (2018) Interpretable adversarial perturbation in input embedding space for text. arXiv preprint 1805.02917. Available at https://arxiv.org/abs/1805.02917.

Schlarmann C and Hein M (2023) On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. New York City, NY: IEEE, pp. 3677–3685.

Schopow N, **Osterhoff G and Baur D** (2023) Applications of the natural language processing tool chatgpt in clinical practice: Comparative study and augmented systematic review. *JMIR Medical Informatics 11*, e48933.

Shaw L (2025) *Sentiment datasets.* Zenodo. https://doi.org/10.5281/zenodo.15190424.

Shaw L and Ekin T (2024) Bertguard: Robust text classification against adversarial attacks. Preprint, 10.22541/au.172556904.43002725/v1. Available at https://www.authorea.com/users/703889/articles/1222619-bertguard-robust-text-classification-against-adversarial-attacks.

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

Shaw L, **Ansari MW and Ekin T** (2025) Adversarial natural language processing: Overview, challenges and future directions. In *Proceedings of 58th Hawaii International Conference on System Sciences (HICSS)*. Honolulu, HI: University of Hawaii Press, p. 904.

Singh K, **Grover SS and Kumar RK** (2022) Cyber security vulnerability detection using natural language processing. In *2022 IEEE World AI IoT Congress (AIIoT)*. New York City, NY: IEEE, pp. 174–178.

Story P, **Zimmeck S**, **Ravichander A**, **Smullen D**, **Wang Z**, **Reidenberg J**, **Cameron Russell N and Sadeh, N.** (2019) Natural language processing for mobile app privacy compliance. In *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*, Vol. *2*. Washington, DC: Association for Advancement of Artificial Intelligence, p. 4.

Strubell E, **Ganesh A and McCallum A** (2020) Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. *34*. Washington, DC: Association for Advancement of Artificial Intelligence, pp. 13693–13696.

Vaswani A, **Shazeer N**, **Parmar N**, **Uszkoreit J**, **Jones L**, **Gomez AN and Polosukhin I** (2017) Attention is all you need. *Advances in Neural Information Processing Systems 30*, 1–11.

Wallace E, **Rodriguez P**, **Feng S**, **Yamada I and 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.

https://doi.org/10.1017/dap.2025.10030 Published online by Cambridge University Press

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

Wang Y and Bansal M (2018) Robust machine comprehension models via adversarial training. arXiv preprint 1804.06473. Available at https://arxiv.org/abs/1804.06473.

Wang B, **Xu C**, **Wang S**, **Gan Z**, **Cheng Y**, **Gao J**, **Awadallah AH and Li B** (2021a) Adversarial glue: A multitask benchmark for robustness evaluation of language models. arXiv preprint 2111.02840. Available at https://arxiv.org/abs/2111.02840.

Wang X, **Hao J**, **Yang Y and He K** (2021b) Natural language adversarial defense through synonym encoding. In *Uncertainty in Artificial Intelligence*. Cambridge, MA: Proceedings of Machine Learning Research, pp. 823–833.

Webster J and Watson RT (2002) Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, xiii–xxiii.

Wong A, **Plasek JM**, **Montecalvo SP and Zhou L** (2018) Natural language processing and its implications for the future of medication safety: A narrative review of recent advances and challenges. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy 38*(8), 822–841.

Wymberry C and 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*. Cham: Springer, pp. 81–116.

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

Yoo JY and Qi Y (2021) Towards improving adversarial training of NLP models. arXiv preprint 2109.00544. Available at https://arxiv.org/abs/2109.00544.

Zang Y, **Qi F**, **Yang C**, **Liu Z**, **Zhang M**, **Liu Q and Sun M** (2019) Word-level textual adversarial attacking as combinatorial optimization. arXiv preprint 1910.12196. Available at https://arxiv.org/abs/1910.12196.

Zeng J, **Li J**, **Song Y**, **Gao C**, **Lyu MR and King I** (2018) Topic memory networks for short text classification. arXiv preprint 1809.03664. Available at https://arxiv.org/abs/1809.03664.

Zhang H, **Zhou H**, **Miao N and Li L** (2020a) Generating fluent adversarial examples for natural languages. arXiv preprint 2007.06174. Available at https://arxiv.org/abs/2007.06174.

Zhang WE, **Sheng QZ**, **Alhazmi A and Li C** (2020b) 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 and Ji Q** (2020) Bayesian adversarial human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York City, NY: IEEE, pp. 6225–6234.

Zou W, **Huang S**, **Xie J**, **Dai X and Chen J** (2019) A reinforced generation of adversarial examples for neural machine translation. arXiv preprint 1911.03677. Available at https://arxiv.org/abs/1911.03677.

Zou W, **Geng R**, **Wang B and Jia J** (2024) Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. arXiv preprint 2402.07867. Available at https://arxiv.org/abs/2402.07867.

Cite this article: Shaw L, Ansari MW and Ekin T (2025). Adversarial natural language processing: overview, challenges, and policy implications. *Data & Policy*, 7: e64. doi:10.1017/dap.2025.10030