CURRENT RESEARCH JOURNAL OF PEDAGOGICS (ISSN: 2767-3278)

VOLUME: Vol.07 Issue04 2026
Page: - 01-13

RESEARCH ARTICLE OPEN ACCESS

Assessing Neural Text Systems Susceptibility to Data Contamination in Limited-Data Environments

Dr. Faisal Al-Nuaimi

Department of Educational Sciences, Qatar University, Doha, Qatar

Received: 08 February 2026 **Accepted:** 05 March 2026 Published**:** 01 April 2026

ABSTRACT

The rapid advancement of neural text systems, particularly large-scale pretrained language models, has transformed natural language processing (NLP) across diverse applications. However, their dependence on vast datasets raises critical concerns regarding vulnerability to data contamination, especially in limited-data environments. This paper investigates the susceptibility of neural text systems to various forms of data poisoning and contamination under constrained data conditions, with a focus on low-resource linguistic contexts. Drawing upon interdisciplinary perspectives from machine learning security, data ethics, and computational linguistics, the study examines how data scarcity amplifies risks associated with adversarial manipulation, bias propagation, and representational distortion.

The research synthesizes existing frameworks of data poisoning attacks, including clean-label, backdoor, and federated poisoning mechanisms, while situating them within the structural limitations of low-resource datasets. Theoretical grounding is established through analyses of model generalization, transfer learning dynamics, and statistical dependency structures inherent in neural architectures. Furthermore, the study explores how limited corpus diversity intensifies model sensitivity to corrupted inputs, leading to systemic degradation in performance, fairness, and robustness.

A conceptual model is developed to illustrate the interaction between dataset quality, model architecture, and adversarial interference. Through analytical evaluation, the study demonstrates that neural text systems operating in low-resource environments exhibit disproportionately higher vulnerability to contamination due to overfitting tendencies, reliance on pretrained representations, and insufficient noise filtering mechanisms. Additionally, the research highlights the role of socio-technical factors, including data curation practices and algorithmic governance, in shaping model resilience.

The findings underscore the necessity for robust data validation protocols, secure training pipelines, and adaptive learning strategies tailored to constrained environments. The paper contributes to ongoing discourse on trustworthy AI by identifying critical vulnerabilities in contemporary NLP systems and proposing strategic directions for enhancing resilience against data contamination. Ultimately, it emphasizes that safeguarding neural text systems requires not only technical innovation but also ethical and institutional interventions.

Keywords: Neural text systems; data contamination; low-resource languages; data poisoning; adversarial attacks; NLP security; pretrained models; robustness; machine learning ethics.

INTRODUCTION

The evolution of neural text systems has redefined computational approaches to language understanding, enabling unprecedented capabilities in tasks such as translation, summarization, and conversational interaction. Central to this transformation is the emergence of large-scale pretrained language models, which leverage

extensive datasets to learn contextual representations of language (Devlin et al., 2019; Brown, 2020; Raffel, 2020). These models operate on the premise that exposure to massive corpora facilitates generalizable linguistic knowledge, which can then be adapted to downstream tasks through fine-tuning or prompting strategies (Han, 2021; Qiu et al., 2020).

Despite these advancements, the reliance on data-intensive training paradigms introduces critical vulnerabilities. Neural text systems are inherently sensitive to the quality and integrity of training data, making them susceptible to various forms of contamination, including noise, bias, and adversarial manipulation (Goldblum, 2023; Pitropakis et al., 2019). While such vulnerabilities exist across all data regimes, they become particularly pronounced in limited-data environments, where the scarcity of training samples amplifies the impact of corrupted inputs.

Low-resource settings, characterized by insufficient annotated data and limited linguistic representation, present unique challenges for NLP systems (Ogueji et al., 2021; Agerri, 2020). In such contexts, models often rely heavily on transfer learning from high-resource languages, which may introduce structural biases and reduce robustness (Artetxe et al., 2022). Furthermore, the limited diversity of training data constrains the model's ability to generalize, increasing susceptibility to overfitting and adversarial exploitation.

Data contamination in neural text systems can manifest in multiple forms. Intentional contamination, such as data poisoning attacks, involves the insertion of malicious samples designed to manipulate model behavior (Shafahi, 2018; Huang et al., 2020). Unintentional contamination, on the other hand, arises from noisy or biased data sources, leading to distorted representations and unfair outcomes (Bender et al., 2021; Hutchinson et al., 2020). Both forms pose significant risks, particularly in applications involving sensitive domains such as healthcare, governance, and social media analysis.

The problem is further complicated by the opaque nature of neural architectures. Unlike traditional rule-based systems, neural models operate as complex, high-dimensional functions, making it difficult to trace the influence of individual data points (Bommasani, 2021). This opacity hinders the detection and mitigation of contamination, allowing vulnerabilities to persist undetected. Moreover, as models scale in size and complexity, their susceptibility to subtle perturbations increases, raising concerns about their reliability and security (Narayanan, 2021).

The relevance of this study is underscored by the growing deployment of NLP systems in real-world applications. From automated content moderation to decision-support systems, neural text models are increasingly integrated into socio-technical infrastructures. In low-resource environments, where data limitations are inherent, the consequences of model failure can be particularly severe, exacerbating existing inequalities and undermining trust in AI systems (Kozyreva et al., 2021).

The primary objective of this research is to systematically assess the susceptibility of neural text systems to data contamination in limited-data environments. The study aims to (1) analyze the structural and functional factors contributing to vulnerability, (2) examine the interaction between data scarcity and adversarial manipulation, and (3) propose conceptual frameworks for enhancing model robustness. By integrating insights from machine learning security and computational linguistics, the research seeks to provide a comprehensive understanding of the challenges and potential solutions.

The scope of the study encompasses both theoretical analysis and conceptual modeling. While empirical experimentation is referenced through existing literature, the primary focus is on synthesizing knowledge to develop a coherent framework for understanding contamination dynamics. This approach allows for a deeper exploration of underlying mechanisms, rather than isolated case studies.

The significance of this research lies in its contribution to the broader discourse on trustworthy AI. As neural text systems continue to evolve, ensuring their reliability and security becomes paramount. By highlighting the vulnerabilities associated with data contamination, particularly in low-resource settings, the study provides critical insights for researchers, practitioners, and policymakers. It emphasizes the need for interdisciplinary approaches that combine technical innovation with ethical considerations.

LITERATURE REVIEW

The study of neural text systems and their vulnerabilities to data contamination is situated at the intersection of

multiple research domains, including machine learning, computational linguistics, and cybersecurity. Existing literature provides a comprehensive foundation for understanding both the capabilities and limitations of these systems, particularly in relation to data dependency and adversarial threats.

Early developments in neural language modeling emphasized the importance of large-scale data in achieving high performance. Foundational works such as transformer architectures (Vaswani et al., 2017) and subsequent models like BERT (Devlin et al., 2019) and GPT (Radford et al., 2018; Brown, 2020) demonstrated that increased data and computational resources lead to improved generalization. These models rely on self-attention mechanisms to capture contextual relationships, enabling them to process complex linguistic structures efficiently (Wolf, 2020).

However, the emphasis on scale has also introduced new challenges. Bender et al. (2021) critically examine the environmental and ethical implications of large language models, highlighting issues related to data bias and representational harm. Similarly, Bommasani (2021) discusses the risks associated with foundation models, including their susceptibility to unintended behaviors arising from training data anomalies.

In the context of low-resource languages, research has focused on adapting pretrained models to environments with limited data availability. Ogueji et al. (2021) demonstrate that multilingual models can achieve reasonable performance even with small datasets, though their effectiveness depends on the quality and relevance of pretrained representations. Agerri (2020) and Etxaniz (2024) further explore the challenges of developing language models for underrepresented languages, emphasizing the importance of tailored data curation strategies.

Artetxe et al. (2022) provide a critical analysis of corpus quality, arguing that the effectiveness of language models is highly dependent on the integrity of training data. Their findings suggest that low-quality data can significantly degrade model performance, particularly in low-resource settings where alternative data sources are scarce. This insight is crucial for understanding the impact of data contamination.

The field of machine learning security offers extensive research on adversarial attacks, including data poisoning. Shafahi (2018) introduces targeted clean-label poisoning attacks, demonstrating how adversaries can manipulate training data without altering labels. Huang et al. (2020) extend this work by proposing general-purpose poisoning methods that can be applied across various models. Kurita et al. (2020) and Li et al. (2021) explore weight poisoning and backdoor attacks, highlighting vulnerabilities in pretrained models.

Goldblum (2023) provides a comprehensive survey of dataset security, categorizing different types of poisoning attacks and their implications. Pitropakis et al. (2019) offer a taxonomy of machine learning attacks, including evasion and poisoning strategies, which serve as a foundational framework for analyzing vulnerabilities.

In addition to intentional attacks, unintentional data contamination has been widely studied. Hutchinson et al. (2020) and Basta et al. (2019) examine biases in NLP models, demonstrating how skewed datasets can lead to discriminatory outcomes. Tan and Celis (2019) further analyze intersectional biases, highlighting the complexity of fairness issues in language models.

Privacy-preserving techniques have also been explored as potential mitigation strategies. Coavoux et al. (2018) propose methods for generating secure text representations, while Maheshwari et al. (2022) introduce differentially private encoders to protect sensitive information. However, these approaches often involve trade-offs between privacy and model performance.

The concept of data contamination is not limited to modern machine learning. Historical analyses of data processing systems, such as those by Agar (2003) and Campbell-Kelly (1990, 1996), reveal that data quality has long been a critical factor in computational systems. These studies provide valuable context for understanding contemporary challenges.

Despite extensive research, several gaps remain. Most studies focus on high-resource environments, with limited attention to the unique vulnerabilities of low-resource settings. Additionally, while individual attack methods have been analyzed, there is a lack of integrated frameworks that consider the interplay between data scarcity, model architecture, and contamination dynamics.

This paper addresses these gaps by synthesizing insights from diverse research areas to develop a comprehensive

understanding of neural text system vulnerabilities in limited-data environments. It builds upon existing literature while introducing new perspectives on the interaction between data quality and model robustness.

MAIN PART

Theoretical Foundations of Data Contamination in Neural Systems

Data contamination in neural text systems can be conceptualized as the introduction of distortions within the training distribution that alter model behavior. From a statistical perspective, machine learning models aim to approximate a function that maps inputs to outputs based on observed data. When the training data is contaminated, the underlying distribution becomes skewed, leading to biased estimations and degraded generalization performance.

Theoretical frameworks from statistical learning theory provide insight into this phenomenon. Models trained on limited datasets are particularly prone to overfitting, as they attempt to capture noise rather than underlying patterns. In such scenarios, even minor contamination can significantly influence model parameters, resulting in disproportionate effects on predictions (Peters et al., 2019).

Transfer learning introduces additional complexity. Pretrained models encode knowledge from large corpora, which may not align with the target domain. When fine-tuned on contaminated low-resource data, these models can amplify existing biases or incorporate malicious patterns, leading to compounded vulnerabilities (Han, 2021).

From a systems perspective, data contamination interacts with model architecture and training dynamics. Neural networks rely on gradient-based optimization, which is sensitive to outliers and adversarial inputs. Poisoned data points can manipulate gradients, steering the model toward undesirable behaviors (Kurita et al., 2020).

Taxonomy of Data Contamination and Poisoning Mechanisms

Data contamination can be broadly categorized into intentional and unintentional forms. Intentional contamination includes adversarial attacks designed to compromise model integrity. These attacks can be further divided into clean-label poisoning, backdoor attacks, and federated poisoning strategies.

Clean-label poisoning involves injecting malicious samples that appear legitimate but are crafted to influence model behavior (Shafahi, 2018). Backdoor attacks introduce hidden triggers that cause the model to produce specific outputs when activated (Li et al., 2021). Federated poisoning targets distributed learning systems, where adversaries manipulate local updates to corrupt the global model (Sun et al., 2022).

Unintentional contamination arises from data collection and preprocessing practices. Noise, annotation errors, and sampling biases can introduce distortions that affect model performance. In low-resource settings, these issues are exacerbated due to limited data availability and reliance on imperfect sources (Artetxe et al., 2022).

The interaction between these contamination types creates complex vulnerability landscapes. For example, biased datasets can make models more susceptible to adversarial attacks, as existing distortions provide entry points for manipulation.

LITERATURE REVIEW

The study of neural text systems and their susceptibility to data contamination in low-resource environments is situated at the intersection of natural language processing (NLP), machine learning security, and data ethics. Existing literature provides a comprehensive foundation for understanding both the capabilities of modern language models and the vulnerabilities that emerge due to data limitations and adversarial manipulation.

A central theme in the literature is the evolution of large-scale language models and their dependence on extensive pre-training. Foundational works such as those by Devlin et al. (2019) and Brown (2020) demonstrate the effectiveness of transformer-based architectures in capturing contextual semantics across diverse linguistic tasks. Similarly, Radford et al. (2018) and Raffel (2020) emphasize the scalability and adaptability of such models through unsupervised and transfer learning paradigms. However, these studies also implicitly highlight a structural dependency on high-quality, large-scale datasets, which becomes problematic in low-resource scenarios.

Research focusing specifically on low-resource languages underscores the challenges associated with limited data availability. Ogueji et al. (2021) and Agerri (2020) illustrate that pre-trained multilingual models can partially mitigate data scarcity, yet their performance is highly sensitive to corpus quality. Artetxe et al. (2022) further argue that dataset quality plays a more critical role than quantity in such settings, as noisy or unrepresentative data can significantly degrade model performance. This establishes a direct link between data integrity and model reliability.

The literature on data contamination and poisoning attacks provides a critical lens for understanding vulnerabilities. Pitropakis et al. (2019) offer a comprehensive taxonomy of machine learning attacks, categorizing them into poisoning, evasion, and inference attacks. Within this framework, Shafahi et al. (2018) and Huang et al. (2020) demonstrate how adversaries can inject malicious samples into training datasets to manipulate model behavior. Kurita et al. (2020) and Li et al. (2021) extend this analysis to pre-trained language models, showing that weight poisoning and backdoor attacks can persist across training stages and remain undetected.

A significant body of work also explores adversarial interactions with transformer-based systems. Misra (2019) and De Wynter (2020) highlight the susceptibility of these models to black-box attacks, where adversaries exploit model outputs without direct access to internal parameters. Pang (2020) further distinguishes between adversarial inputs and poisoned models, emphasizing that both forms of attack can coexist and reinforce each other, particularly in systems with limited data diversity.

Bias and fairness constitute another critical dimension of the literature. Studies by Basta et al. (2019) and Tan and Celis (2019) reveal that contextualized word embeddings often encode gender and social biases present in training data. Hutchinson et al. (2020) extend this analysis to accessibility concerns, demonstrating that NLP systems can inadvertently marginalize individuals with disabilities. In low-resource environments, where datasets may lack diversity, these biases are not only preserved but amplified, raising significant ethical concerns.

The concept of foundation models, as discussed by Bommasani (2021), provides a broader theoretical framework for understanding both the opportunities and risks associated with large-scale pre-trained systems. Bender et al. (2021) critically examine the societal implications of such models, warning against over-reliance on data-driven approaches without sufficient consideration of data provenance and quality. These perspectives are particularly relevant in the context of data contamination, as they highlight the systemic risks associated with opaque training processes.

In addition to adversarial and ethical considerations, the literature also addresses privacy and security mechanisms. Coavoux et al. (2018) and Maheshwari et al. (2022) propose privacy-preserving techniques such as differential privacy and secure embeddings to mitigate data leakage and contamination risks. However, these approaches often involve trade-offs between model performance and security, particularly in low-resource settings where data is already scarce.

Another important strand of research examines the scalability and infrastructure of neural text systems. Narayanan (2021) discusses large-scale training frameworks that enable the development of trillion-parameter models, while Han (2021) provides a comprehensive overview of the evolution of pre-trained models. Although these advancements enhance model capabilities, they also increase the attack surface, making it more challenging to detect and mitigate contamination at scale.

Despite the extensive body of work, several research gaps remain evident. First, while individual studies address either low-resource challenges or data poisoning vulnerabilities, there is limited integration of these perspectives. The interaction between data scarcity and adversarial contamination remains underexplored, particularly in terms of how structural limitations amplify attack effectiveness. Second, existing evaluation metrics often fail to capture the nuanced impact of contamination on fairness and robustness, indicating a need for more comprehensive assessment frameworks. Third, there is a lack of empirical studies focusing on real-world low-resource environments, where data quality and availability vary significantly.

In summary, the literature establishes a strong foundation for understanding the capabilities and vulnerabilities of neural text systems. It highlights the critical role of data quality, the diverse range of adversarial threats, and the ethical implications of biased and contaminated datasets. However, the intersection of these factors in low-resource

environments remains insufficiently addressed, underscoring the need for further research that integrates technical, theoretical, and ethical perspectives.

METHOD

The core analysis of this study is structured around the interaction between neural text system design, data constraints, and contamination vulnerabilities. This section develops a comprehensive framework that integrates architectural, statistical, and adversarial dimensions to explain how susceptibility emerges and propagates in limited-data environments. The discussion is grounded in existing theoretical models of machine learning, while extending them to account for the unique challenges posed by low-resource settings.

Foundations of Neural Text Systems in Low-Resource Environments

Neural text systems are fundamentally built upon deep learning architectures that model linguistic patterns through distributed representations. Transformer-based models, characterized by self-attention mechanisms, have become the dominant paradigm due to their ability to capture long-range dependencies and contextual semantics (Devlin et al., 2019; Wolf, 2020). However, their effectiveness is closely tied to the availability of large-scale, high-quality datasets.

In low-resource environments, this dependency creates a structural imbalance. Models rely heavily on pre-trained representations derived from high-resource languages or domains, which may not align with the linguistic and cultural characteristics of the target data. This mismatch leads to representational gaps, where certain linguistic features are underrepresented or misinterpreted (Agerri, 2020; Ogueji et al., 2021). As a result, the model's internal representation space becomes uneven, increasing its sensitivity to noise and contamination.

From a theoretical standpoint, this phenomenon can be understood through the lens of statistical learning theory. When training data is limited, the model's ability to approximate the true data distribution is constrained, leading to higher variance and overfitting. In such conditions, even minor perturbations in the training data can significantly influence model parameters (Artetxe et al., 2022). This sensitivity forms the basis of vulnerability to data contamination.

Moreover, transfer learning, while beneficial for mitigating data scarcity, introduces additional complexities. Pre-trained models encapsulate patterns learned from large corpora, which may include biases, noise, or adversarial artifacts. During fine-tuning, these inherited characteristics interact with the limited target data, often dominating the learning process (Qiu et al., 2020; Radford et al., 2018). Consequently, the model's behavior reflects a combination of pre-training biases and local data limitations.

Data Contamination: Types, Sources, and Mechanisms

Data contamination in neural text systems encompasses a broad spectrum of phenomena, ranging from unintentional noise to deliberate adversarial manipulation. Understanding the types and sources of contamination is essential for analyzing their impact on model performance and reliability.

One primary category is label contamination, where incorrect or misleading labels are assigned to training samples. This can occur due to human annotation errors, automated labeling processes, or adversarial interventions. Label contamination distorts the mapping between input features and output classes, leading to erroneous decision boundaries (Xiao et al., 2012). In low-resource settings, the impact is particularly severe because each labeled instance carries significant weight in the training process.

Another critical category is feature-level contamination, often associated with data poisoning attacks. In such cases, adversaries inject carefully crafted samples into the training dataset to manipulate model behavior. These samples may appear benign but contain subtle patterns designed to influence model predictions (Shafahi et al., 2018; Huang et al., 2020). Clean-label poisoning, for example, preserves correct labels while embedding adversarial features, making detection challenging.

Backdoor attacks represent a more sophisticated form of contamination. These attacks introduce hidden triggers into the training data, enabling the model to produce specific outputs when the trigger is present (Li et al., 2021; Kurita et al., 2020). In low-resource environments, the limited diversity of training data increases the likelihood that such triggers will be learned and retained by the model.

Unintentional contamination also plays a significant role. Noisy data, domain mismatches, and biased corpora can all introduce distortions in model learning. For instance,

datasets collected from online sources may contain duplicated, inconsistent, or contextually irrelevant information, which affects the quality of learned representations (Ahmad et al., 2023). In low-resource scenarios, the lack of alternative data sources prevents effective filtering or correction.

Additionally, contamination can arise from data preprocessing and augmentation techniques. While these methods are intended to enhance data diversity, they may inadvertently introduce artifacts or amplify existing biases. For example, synthetic data generation techniques can replicate noise patterns present in the original dataset, thereby reinforcing contamination rather than mitigating it.

Impact of Limited Data on Model Robustness and Generalization

The relationship between data availability and model robustness is a central concern in the study of neural text systems. Limited data not only constrains model performance but also amplifies vulnerabilities to contamination and adversarial manipulation.

In terms of generalization, models trained on small datasets often struggle to capture the full variability of the target domain. This leads to overfitting, where the model memorizes training samples rather than learning generalized patterns. Overfitting increases sensitivity to noise and adversarial perturbations, as the model lacks the flexibility to adapt to unseen variations (Peters et al., 2019).

Robustness is similarly affected. In high-resource settings, the presence of diverse training samples provides a form of statistical resilience, enabling the model to ignore outliers and anomalous patterns. In contrast, low-resource models lack this redundancy, making them more susceptible to the influence of contaminated data points. Even a small number of poisoned samples can disproportionately affect model behavior (Goldblum, 2023).

Another critical aspect is the interaction between model scale and data availability. Large models, characterized by millions or billions of parameters, require substantial data to achieve stable training. When applied to low-resource tasks, these models may exhibit unstable behavior, including overfitting and sensitivity to initialization (Han, 2021; Bommasani, 2021). This instability creates opportunities for adversarial exploitation.

The impact of limited data is also evident in bias amplification. When training data lacks diversity, the model tends to reinforce dominant patterns while neglecting minority representations. This leads to skewed outputs and reduced fairness, particularly in applications involving sensitive attributes such as gender, ethnicity, or disability (Basta et al., 2019; Hutchinson et al., 2020). In extreme cases, biased outputs can be interpreted as a form of systemic contamination.

Furthermore, evaluation challenges arise in low-resource environments. Standard validation techniques rely on sufficiently large and representative test sets, which may not be available. As a result, model performance metrics may not accurately reflect real-world behavior, masking the effects of contamination and reducing the reliability of evaluation processes.

In summary, limited data fundamentally alters the dynamics of neural text system training and evaluation. It increases model sensitivity, reduces robustness, and amplifies the impact of contamination, thereby creating a complex landscape of vulnerabilities that require specialized mitigation strategies.

Technical Architecture of Neural Text Systems under Data Constraints

The technical architecture of neural text systems operating in limited-data environments is fundamentally shaped by the interplay between pre-training, fine-tuning, and data ingestion pipelines. Modern architectures such as transformer-based models rely on large-scale pre-training followed by domain-specific adaptation (Devlin et al., 2019; Raffel, 2020). However, in low-resource settings, the architecture must compensate for insufficient data diversity and volume, which introduces structural vulnerabilities to contamination.

At the core of these systems lies a multi-layer transformer architecture composed of self-attention mechanisms that encode contextual relationships across tokens (Vaswani et al., 2017; Wolf, 2020). In limited-data environments, the reliance on pre-trained representations becomes disproportionately high, as downstream fine-tuning datasets are often insufficient to override biases or injected perturbations. Consequently, the architecture amplifies inherited noise or malicious signals embedded within training corpora.

The data pipeline typically includes data collection, preprocessing, tokenization, embedding generation, and model training. Each stage presents a potential point of contamination. For instance, tokenization schemes in multilingual models may inadequately represent low-resource languages, leading to fragmented semantic encoding and increased susceptibility to adversarial manipulation (Pires et al., 2019; Wang et al., 2020). Similarly, preprocessing steps such as normalization and filtering can inadvertently preserve adversarial patterns if they are statistically subtle.

From a systems perspective, distributed training frameworks such as Megatron introduce additional vulnerabilities by decentralizing data handling and gradient updates (Narayanan, 2021). In such architectures, poisoned data can propagate across nodes, contaminating model weights at scale. This is particularly problematic in federated or collaborative training environments where data provenance is difficult to verify (Sun et al., 2022; Tolpegin et al., 2020).

Moreover, parameter-efficient fine-tuning methods, while beneficial for low-resource adaptation, may inadvertently localize vulnerabilities. Techniques such as adapter layers or prompt tuning concentrate learning in specific parameter subsets, making them more sensitive to targeted poisoning attacks (Peters et al., 2019). Thus, architectural efficiency introduces a trade-off between adaptability and robustness.

Classification Mechanisms and Vulnerability Propagation

Classification mechanisms in neural text systems are intrinsically dependent on learned representations derived from training data. In low-resource environments, these mechanisms often rely on transfer learning, where pre-trained embeddings are adapted to specific classification tasks such as sentiment analysis, named entity recognition, or propaganda detection (Ahmad et al., 2023; Petroni et al., 2019).

The fundamental issue arises from the alignment between representation space and decision boundaries. When training data is contaminated, either through label manipulation or semantic distortion, the learned decision boundaries become skewed. This leads to systematic misclassification, particularly for underrepresented classes. For example, in low-resource languages, even minor perturbations in training data can disproportionately affect classification outcomes due to limited redundancy (Ogueji et al., 2021; Artetxe et al., 2022).

Backdoor and poisoning attacks exploit this sensitivity by embedding hidden triggers within training data. These triggers remain dormant during normal operation but activate under specific conditions, causing predictable misclassifications (Li et al., 2021; Kurita et al., 2020). The effectiveness of such attacks is amplified in low-resource settings, where the model lacks sufficient examples to generalize beyond manipulated patterns.

Furthermore, bias propagation within classification mechanisms is exacerbated by data scarcity. Studies on gender and social biases demonstrate that neural representations encode and reproduce societal biases present in training data (Basta et al., 2019; Hutchinson et al., 2020). In low-resource contexts, the absence of corrective data leads to the reinforcement of these biases, which can be interpreted as a form of unintentional data contamination.

Adversarial inputs also challenge classification robustness by exploiting model overconfidence. Black-box attacks generate inputs that appear semantically valid but are strategically crafted to induce misclassification (Misra, 2019; De Wynter, 2020). In low-resource systems, the lack of robust decision boundaries increases susceptibility to such attacks, highlighting the need for more resilient classification frameworks.

Analytical Evaluation of Contamination Effects

The evaluation of contamination effects in neural text systems requires a multidimensional analytical framework that considers accuracy degradation, robustness, fairness, and interpretability. Traditional evaluation metrics such as accuracy and F1-score are insufficient to capture the nuanced impact of data poisoning, particularly in low-resource environments.

Empirical studies indicate that even small proportions of poisoned data can lead to significant performance degradation (Goldblum, 2023; Shafahi et al., 2018). This effect is magnified in limited-data settings, where each data point carries greater influence on model learning. Consequently, contamination introduces both localized errors and systemic distortions in model behavior.

A critical dimension of evaluation is the distinction

between clean-label and dirty-label poisoning. Clean-label attacks maintain correct labels while subtly altering input features, making detection challenging (Huang et al., 2020). In contrast, dirty-label attacks involve explicit label manipulation, which can be more easily identified but equally damaging (Xiao et al., 2012). Both forms have distinct implications for model reliability and require different mitigation strategies.

Another important aspect is the persistence of contamination across training iterations. Once embedded in model weights, poisoned patterns can persist even after retraining or fine-tuning, particularly when training data remains limited (Pang, 2020). This persistence underscores the need for proactive data validation and robust training protocols.

The interaction between contamination and model scale also warrants attention. Larger models exhibit greater capacity to memorize and propagate poisoned patterns, as highlighted in studies on foundation models (Bommasani, 2021; Bender et al., 2021). While scaling improves generalization, it simultaneously increases the risk of amplifying hidden vulnerabilities.

Finally, the evaluation of contamination must consider ethical and societal implications. Data poisoning can lead to biased or harmful outputs, undermining trust in AI systems and exacerbating existing inequalities (Kozyreva et al., 2021). In low-resource settings, where data often reflects marginalized communities, the consequences of contamination are particularly severe.

RESULTS

The analysis reveals that neural text systems operating in limited-data environments exhibit a heightened susceptibility to data contamination due to structural, statistical, and representational constraints. The findings indicate that data scarcity significantly amplifies the influence of individual data points, thereby increasing the effectiveness of both intentional poisoning attacks and unintentional biases embedded within training corpora.

First, the evaluation demonstrates that contamination disproportionately affects classification accuracy in low-resource contexts. Models trained on limited datasets show rapid degradation in performance even when exposed to minimal levels of poisoned data. This is attributed to the lack of redundancy and diversity in training samples, which prevents the model from generalizing beyond manipulated patterns (Artetxe et al., 2022; Ogueji et al., 2021).

Second, the study identifies a strong correlation between model architecture and vulnerability. Transformer-based systems, while highly expressive, are particularly prone to memorizing adversarial patterns due to their attention mechanisms and large parameter spaces (Devlin et al., 2019; Han, 2021). This memorization effect enables the persistence of backdoor triggers and poisoned representations across training cycles.

Third, the findings highlight the role of pre-training in shaping susceptibility. Models initialized with large-scale pre-trained representations inherit both strengths and weaknesses from their training data. In low-resource fine-tuning scenarios, these inherited characteristics dominate model behavior, making it difficult to correct for contamination introduced during pre-training (Qiu et al., 2020; Radford et al., 2018).

Fourth, the analysis reveals that classification mechanisms are particularly vulnerable to bias amplification. Contaminated data leads to skewed decision boundaries, resulting in systematic misclassification of underrepresented classes. This effect is further intensified by the absence of sufficient corrective data, leading to persistent bias in model outputs (Basta et al., 2019; Hutchinson et al., 2020).

Finally, the results underscore the complexity of detecting and mitigating contamination. Clean-label poisoning and subtle adversarial manipulations often evade traditional detection methods, necessitating more sophisticated evaluation frameworks and defense strategies (Goldblum, 2023; Kurita et al., 2020).

DISCUSSION

The findings provide critical insights into the interplay between data scarcity and model vulnerability, emphasizing the need for a paradigm shift in the design and evaluation of neural text systems. The observed susceptibility to contamination is not merely a technical limitation but reflects deeper structural issues related to data representation, model architecture, and training methodologies.

From a theoretical perspective, the results align with the

concept of statistical fragility in low-resource learning. When training data is limited, the model's reliance on prior knowledge and individual data points increases, creating a feedback loop that amplifies both useful patterns and harmful perturbations (Bommasani, 2021; Bender et al., 2021). This highlights the importance of understanding data quality as a central determinant of model robustness.

Practically, the findings suggest that existing approaches to model training and evaluation are insufficient for low-resource environments. Standard practices such as random data splitting and cross-validation fail to account for the disproportionate impact of contaminated samples. Instead, there is a need for targeted validation strategies that explicitly assess robustness to adversarial and biased inputs (Pitropakis et al., 2019).

The discussion also reveals significant trade-offs between model performance and security. Techniques that enhance model capacity, such as scaling and pre-training, simultaneously increase vulnerability to poisoning attacks. Similarly, methods designed to improve efficiency, such as parameter-efficient fine-tuning, may inadvertently concentrate vulnerabilities in specific components of the model (Peters et al., 2019).

Another critical implication concerns ethical and societal dimensions. The amplification of biases in low-resource settings raises concerns about fairness and inclusivity, particularly for marginalized linguistic communities. Contaminated models may produce outputs that reinforce stereotypes or propagate misinformation, thereby undermining the broader goals of equitable AI development (Kozyreva et al., 2021).

Despite these insights, the study acknowledges several limitations. The analysis is primarily theoretical and based on existing literature, which may not fully capture the diversity of real-world scenarios. Additionally, the rapidly evolving nature of neural architectures and attack strategies necessitates continuous reassessment of vulnerabilities and defenses.

CONCLUSION

This research provides a comprehensive examination of the susceptibility of neural text systems to data contamination in limited-data environments. By integrating theoretical analysis with insights from existing literature, the study highlights the structural, statistical, and ethical dimensions of vulnerability in low-resource settings.

The findings demonstrate that data scarcity significantly amplifies the impact of contamination, affecting model accuracy, robustness, and fairness. Transformer-based architectures, while powerful, exhibit inherent vulnerabilities due to their capacity to memorize and propagate adversarial patterns. Classification mechanisms further exacerbate these issues by reinforcing biases and distorted decision boundaries.

The study contributes to the field by emphasizing the need for robust data curation, advanced evaluation frameworks, and secure training methodologies. It also underscores the importance of addressing ethical considerations, particularly in the context of marginalized languages and communities.

Future research should focus on developing adaptive defense mechanisms, improving data quality assessment techniques, and exploring alternative architectures that balance performance with robustness. Additionally, empirical validation of theoretical findings in real-world low-resource scenarios remains a critical area for further investigation.

REFERENCES

1. J. Agar, The Government Machine: A Revolutionary History of the Computer. Cambridge, MA, USA : MIT Press, 2003.

2. P. N. Ahmad, Y. Liu, G. Ali, M. A. Wani, and M. ElAffendi, "Robust benchmark for propagandist text detection and mining high-quality data," Mathematics, vol. 11, no. 12, 2023, Art. no. 2668.

3. R. Agerri, "Give your text representation models some love: The case for basque," in Proc. 12th Lang. Resour. Eval. Conf., Marseille, France, May 2020, pp. 4781–4788. [Online]. Available: https://aclanthology.org/2020.lrec-1.588

4. S. A. Aluko, "How many Nigerians? An analysis of Nigeria's census problems, 1901-63," J. Modern Afr. Stud., vol. 3, no. 3, pp. 371–392, 1965, doi: 10.1017/S0022278X00006170.

5. M. Artetxe, I. Aldabe, R. Agerri, O. Perez-De-Viñaspre, and A. Soroa, "Does corpus quality really

matter for low-resource languages?," in Proc. 2022 Conf. Empirical Methods Natural Lang. Process., Dec. 2022, pp. 7383–7390. [Online]. Available: https://aclanthology.org/2022.emnlp-main.499

6. C. Basta, M. R. Costa-jussà, and N. Casas, "Evaluating the underlying gender bias in contextualized word embeddings," in Proc. 1st Workshop Gender Bias Natural Lang. Process., Florence, Italy, Aug. 2019, pp. 33–39. [Online]. Available: https://www.aclweb.org/anthology/W19-3805

7. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in Proc. 2021 ACM Conf. Fairness, Accountability, Transparency, Mar. 2021, pp. 610–623, doi: 10.1145/3442188.3445922.

8. R. Bommasani, "On the opportunities and risks of foundation models," Aug. 2021, arXiv:2108.07258.

9. T. Brown, "Language models are few-shot learners," in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

10. M. Campbell-Kelly, "Information technology and organizational change in the British census, 1801-1911," Inf. Syst. Res., vol. 7, no. 1, pp. 22–36, 1996, doi: 10.1287/isre.7.1.22.

11. M. Campbell-Kelly, "Punched-card machinery," in Computing Before Computers, W. Aspray, Ed., Ames, IA, USA : Iowa State Univ. Press, 1990, pp. 122–155.

12. M. Coavoux, S. Narayan, and S. B. Cohen, "Privacy-preserving neural representations of text," in Proc. 2018 Conf. Empirical Methods Natural Lang. Process., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1–10. [Online]. Available: https://aclanthology.org/D18-1001

13. J. W. Cortada, Before the Computer. IBM, NCR, Burroughs, & Remington Rand & The Industry They Created 1865-1956. Princeton, NJ, USA : Princeton Univ. Press, 1993.

14. A. De Wynter, "Mischief: A simple black-box attack against transformer architectures," Oct. 2020, arXiv:2010.08542.

15. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., Jun. 2019, pp. 4171–4186. [Online]. Available: 10/ggbwf6

16. D. Edgerton, "From innovation to use: Ten eclectic theses on the historiography of technology," Hist. Technol., vol. 16, no. 2, pp. 111–136, 1999, doi: 10.1080/07341519908581961.

17. J. Etxaniz, "Latxa: An open language model and evaluation suite for basque," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics, Aug. 2024, pp. 14952–14972. [Online]. Available: https://aclanthology.org/2024.acl-long.799

18. M. Goldblum, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 2, pp. 1563–1580, Feb. 2023.

19. X. Han, "Pre-trained models: Past, present and future," AI Open, vol. 2, pp. 225–250, Aug. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000231

20. W. R. Huang, J. Geiping, L. Fowl, G. Taylor, and T. Goldstein, "MetaPoison: Practical general-purpose clean-label data poisoning," in Proc. Adv. Neural Inf. Process. Syst., 2020, vol. 33, pp. 12080–12091. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/8ce6fc704072e351679ac97d4a985574-Abstract.html

21. B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl, "Social biases in NLP models as barriers for persons with disabilities," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2020, pp. 5491–5501. [Online]. Available: https://aclanthology.org/2020.acl-main.487

22. P. Kaghazgaran, M. Alfifi, and J. Caverlee, "Wide-ranging review manipulation attacks: Model,

empirical study, and countermeasures," in Proc. 28th ACM Int. Conf. Inf. Knowl. Manage., Nov. 2019, pp. 981–990, doi: 10.1145/3357384.3358034.

23. A. Kozyreva, P. Lorenz-Spreen, R. Hertwig, S. Lewandowsky, and S. M. Herzog, "Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, great britain, and the United States," Humanities Social Sci. Commun., vol. 8, no. 1, pp. 1–11, May 2021. [Online]. Available: 10/gmgpfd

24. K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pretrained models," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Assoc. Comput. Linguistics, Apr. 2020, pp. 2793–2806.

25. L. Li, D. Song, X. Li, J. Zeng, R. Ma, and X. Qiu, "Backdoor attacks on pre-trained models by layerwise weight poisoning," in Proc. 2021 Conf. Empirical Methods Natural Lang. Process., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3023–3032. [Online]. Available: https://aclanthology.org/2021.emnlp-main.241

26. G. Maheshwari, P. Denis, M. Keller, and A. Bellet, "Fair NLP models with differentially private text encoders," in Proc. Find. Assoc. Comput. Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 2022, pp. 6913–6930.

27. V. Misra, "Black box attacks on transformer language models," in Proc. ICLR 2019 Debugging Mach. Learn. Models Workshop, 2019, pp. 1–5.

28. D. Narayanan, "Scaling language model training to a trillion parameters using megatron," Nvidia, Apr. 2021. [Online]. Available: https://developer.nvidia.com/blog/scaling-language-model-training-to-a-trillion-parameters-using-megatron/

29. K. Ogueji, Y. Zhu, and J. Lin, "Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages," in Proc. 1st Workshop Multilingual Representation Learn., Nov. 2021, pp. 116–126. [Online]. Available: https://aclanthology.org/2021.mrl-1.11

30. R. Pang, "A tale of evil twins: Adversarial inputs versus poisoned models," in Proc. 2020 ACM SIGSAC Conf. Comput. Commun. Secur., New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 85–99, doi: 10.1145/3372297.3417253.

31. P. Papadopoulos, O. T. V. Essen, N. Pitropakis, C. Chrysoulas, A. Mylonas, and W. J. Buchanan, "Launching adversarial attacks against network intrusion detection systems for IoT," J. Cybersecurity Privacy, vol. 1, no. 2, pp. 252–273, Jun. 2021. [Online]. Available: https://www.mdpi.com/2624-800X/1/2/14

32. M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? Adapting pretrained representations to diverse tasks," in Proc. 4th Workshop Representation Learn. NLP, Aug. 2019, pp. 7–14.

33. N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A taxonomy and survey of attacks against machine learning," Comput. Sci. Rev., vol. 34, Nov. 2019, Art. no. 100199. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013718303289

34. F. Petroni, "Language models as knowledge bases?," in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., Hong Kong, China, Nov. 2019, pp. 2463–2473. [Online]. Available: https://aclanthology.org/D19-1250

35. X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," Sci. China Technol. Sci., vol. 63, no. 10, pp. 1872–1897, Oct. 2020.

36. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, Tech. Rep., 2018.

37. C. Raffel, "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: https://jmlr.org/papers/volume21/20-074/20-074.pdf

38. R. Schuster, C. Song, E. Tromer, and V. Shmatikov,

"You autocomplete me: Poisoning vulnerabilities in neural code completion," in Proc. 30th USENIX Secur. Symp. USENIX Assoc., 2021, pp. 1559–1575. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/schuster

39. A. Shafahi, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in Proc. 32nd Int. Conf. Neural Inf. Process. Syst., 2018, pp. 6106–6116. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html

40. A. M. Shah and N. Schweiggart, "# Boycottmurree campaign on twitter: Monitoring public response to the negative destination events during a crisis," Int. J. Disaster Risk Reduction, vol. 92, 2023, Art. no. 103734.

41. A. Srivastava, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," in Proc. Trans. Mach. Learn. Res., Jun. 2023. [Online]. Available: https://openreview.net/forum?id=uyTL5Bvosj

42. G. Sun, Y. Cong, J. Dong, Q. Wang, and J. Liu, "Data poisoning attacks on federated machine learning," IEEE Internet Things J., vol. 9, no. 13, pp. 11365–11375, Jul. 2022, arXiv:2004.10020. [Online]. Available: https://ieeexplore.ieee.org/document/9618642

43. Y. C. Tan and L. E. Celis, "Assessing social and intersectional biases in contextualized word representations," in Proc. Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, 2019, pp. 13209–13220. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html

44. V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in Proc. Comput. Secur.–ESORICS 2020, Cham, Switzerland: Springer, Jul. 2020, pp. 480–501.

45. T. Wolf, "Transformers: State-of-the-art natural language processing," in Proc. 2020 Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations, Oct. 2020, pp. 38–45.