Sampling Enhanced Differential Privacy for Large Language Models

Jigyas Sharma
EECS
University of Kansas
Lawrence, Kansas
jigyas@icloud.com

***Abstract*—Fill the abstract after work finishes**
***Index Terms*—Keywords relevant to the paper.**

I. Introduction

Sentiment analysis, a cornerstone of natural language processing (NLP), involves determining the emotional tone behind a piece of text. It has widespread applications in areas such as market analysis, customer service, and social media monitoring. Large language models (LLMs), such as BERT, have achieved state-of-the-art performance in sentiment classification tasks by leveraging their deep contextual understanding of language.

However, as these models become more integrated into decision-making processes, their vulnerability to adversarial attacks has raised significant concerns. Adversarial attacks can subtly alter the input text to mislead the model, resulting in incorrect sentiment classification (e.g., flipping a positive sentiment to a negative one or vice versa). These attacks not only compromise the reliability of the model but also pose risks when the models are deployed in sensitive or high-stakes environments.

Given the increasing reliance on sentiment analysis powered by LLMs, there is a pressing need to develop robust defense mechanisms that can detect and mitigate adversarial attacks. This project aims to explore the vulnerabilities of BERT in sentiment classification and propose a defense strategy that preserves the model's accuracy and reliability, even in the presence of adversarial inputs.

II. Related Work

The vulnerability of machine learning models, particularly large language models (LLMs), to adversarial attacks has been a topic of intense research. Adversarial attacks in NLP often involve minimal perturbations to the input text that are imperceptible to humans but can lead to significant changes in model predictions. This section reviews key works that have contributed to understanding and mitigating these vulnerabilities in the context of sentiment analysis and beyond.

A. Adversarial Attacks on Large Language Models

Goodfellow et al. [3] initially demonstrated the susceptibility of deep learning models to adversarial examples in the context of image classification. Following this, researchers like Ebrahimi et al. [2] and Jin et al. [4] extended this concept to NLP, showing that even state-of-the-art models like BERT can be easily fooled by strategically crafted adversarial inputs. Techniques such as text paraphrasing, synonym substitution, and character-level perturbations have been explored to create adversarial examples that can manipulate sentiment classification models.

Ebrahimi et al. [2] proposed HotFlip, an efficient method for generating adversarial examples by flipping characters in the input text. Jin et al. [4] introduced TextFooler, an attack that substitutes words with their synonyms to generate adversarial examples that are semantically similar to the original text but result in misclassification. These studies have underscored the fragility of LLMs in NLP tasks, especially sentiment analysis.

B. Defense against Adversarial Attacks

To counter adversarial attacks, various defense strategies have been proposed. Adversarial training, introduced by Madry et al. [5], is one of the most prominent techniques, where the model is trained on adversarial examples to improve its robustness. However, adversarial training in NLP is computationally expensive and can lead to overfitting to specific types of attacks.

Another line of defense is based on detecting adversarial examples before they are fed into the model. Wang et al. [7] proposed an approach that uses auxiliary models to detect and filter adversarial inputs in text classification tasks. Similarly, Michel et al. [6] explored the use of gradient-based techniques to identify and discard adversarial examples.

Differential privacy (DP) has also been explored as a defense mechanism in NLP, particularly in scenarios where model interpretability could leak sensitive information. Abadi et al. [1] applied DP to neural networks to ensure that the models generalize well without memorizing specific training data points, thereby reducing vulnerability to certain attacks. However, the application of DP in the context of defending against adversarial attacks on LLMs remains underexplored.

C. Adversarial Attacks and Defenses in Sentiment Analysis

Sentiment analysis, given its wide-ranging applications, has been a focal point in adversarial attack research. Zhang et al. [8] conducted a comprehensive study on the robustness

of sentiment analysis models against adversarial attacks, revealing significant vulnerabilities in models such as BERT. They explored different attack strategies and evaluated the effectiveness of various defense mechanisms.

In terms of defenses, adversarial training has been the most extensively studied approach. However, its application to sentiment analysis is challenging due to the nuanced nature of language and the complexity of capturing semantic meaning while generating adversarial examples. Therefore, there is a growing interest in developing more sophisticated defense mechanisms that can be seamlessly integrated into LLMs like BERT.

III. Methodology

This research aims to explore the vulnerabilities of large language models (LLMs) like BERT in sentiment classification tasks, specifically focusing on adversarial attacks that can manipulate sentiment predictions. The following baseline steps outline the proposed methodology, with the understanding that these steps may evolve as the research advances.

A. Step 1: Exploration of Adversarial Attack Techniques

The initial phase of the project will involve an extensive review and implementation of existing adversarial attack techniques on BERT-based sentiment classification models. We will focus on methods such as:

- **TextFooler** [4]: A word-level attack that substitutes words with synonyms to alter the sentiment classification.
- **HotFlip** [2]: A character-level attack that flips characters in the input text to change the model's output.
- **Semantically Similar Attacks**: Exploring other semantically-preserving attacks that maintain the original meaning while causing misclassification.

The goal of this step is to identify the types of adversarial inputs that most effectively deceive BERT in sentiment classification tasks, and to create a benchmark for evaluating defense mechanisms.

B. Step 2: Development of an Adversarial Attack Targeting BERT

Building on the findings from the first step, we will develop a novel adversarial attack specifically tailored to exploit the weaknesses identified in BERT's sentiment analysis capabilities. This attack will be designed to:

- Target specific vulnerabilities in BERT's architecture, such as its tokenization process or attention mechanism.
- Be subtle enough to evade detection by standard adversarial defense techniques.
- Focus on altering the sentiment classification from positive to negative or vice versa.

This step aims to push the boundaries of current adversarial attack methods, providing a challenging test case for the subsequent defense strategies.

C. Step 3: Design and Implementation of Defense Mechanisms

The third phase of the research will focus on designing and implementing defense mechanisms to protect BERT against the developed adversarial attacks. We will explore the following approaches:

- **Adversarial Training**: Incorporating adversarial examples into the training process to improve model robustness.
- **Differential Privacy (DP)**: Applying DP techniques to the model's training data or explanation outputs to obscure sensitive information and mitigate the effectiveness of adversarial attacks.
- **Detection Mechanisms**: Developing algorithms that can detect and filter adversarial inputs before they are processed by the model.

The effectiveness of these defenses will be evaluated in terms of their ability to maintain model accuracy while reducing the success rate of adversarial attacks.

D. Step 4: Evaluation and Analysis

The final step involves a comprehensive evaluation of the proposed attack and defense mechanisms. This will include:

- **Performance Metrics**: Assessing the accuracy, precision, recall, and F1-score of BERT on clean and adversarially perturbed datasets.
- **Robustness Metrics**: Measuring the decrease in the success rate of adversarial attacks after applying the proposed defenses.
- **Utility vs. Privacy Trade-off**: Analyzing the trade-off between maintaining model utility and ensuring privacy through differential privacy techniques.

This evaluation will provide insights into the effectiveness of the proposed methods and guide further research and development in this area.

E. Flexibility and Iterative Refinement

Given the exploratory nature of this research, the above steps are intended as a baseline framework. The specific methodologies and techniques will be refined iteratively based on experimental findings and emerging insights throughout the course of the project.

IV. Results

Present the results obtained from the research in this section.

V. Conclusion

Summarize the key findings and contributions of the paper.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Han Wang, for her unwavering support and guidance throughout the course of this research. Her insightful explanations of complex concepts and her constant encouragement have been invaluable to the development of this project. Dr. Wang's expertise and mentorship have greatly enriched my understanding of the subject matter, and her dedication to my

academic and professional growth is deeply appreciated. This work would not have been possible without her continued support and belief in my potential.

REFERENCES

[1] Martin Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS'16. ACM, Oct. 2016. DOI: 10.1145/2976749.2978318. URL: http://dx.doi.org/10.1145/2976749.2978318.

[2] Javid Ebrahimi et al. *HotFlip: White-Box Adversarial Examples for Text Classification*. 2018. arXiv: 1712.06751 [cs.CL]. URL: https://arxiv.org/abs/1712.06751.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML]. URL: https://arxiv.org/abs/1412.6572.

[4] Di Jin et al. *Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment*. 2020. arXiv: 1907.11932 [cs.CL]. URL: https://arxiv.org/abs/1907.11932.

[5] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706.06083 [stat.ML]. URL: https://arxiv.org/abs/1706.06083.

[6] Paul Michel et al. *On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models*. 2019. arXiv: 1903.06620 [cs.CL]. URL: https://arxiv.org/abs/1903.06620.

[7] Huaxia Wang and Chun-Nam Yu. *A Direct Approach to Robust Deep Learning Using Adversarial Networks*. 2019. arXiv: 1905.09591 [cs.CV]. URL: https://arxiv.org/abs/1905.09591.

[8] Huan Zhang et al. *Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations*. 2021. arXiv: 2003.08938 [cs.LG]. URL: https://arxiv.org/abs/2003.08938.