CS616 Final Project Report
Adversarial Attacks on Text Classification Models

Farzan Mirza

June 2024

1 Background

Adversarial attacks in natural language processing (NLP) subtly manipulate input to deceive models while preserving human readability. Significantly altering the model's predicted class in **text classification**, while potentially transforming logical relationships, such as converting contradictions into entailments, in **textual entailment**. Due to its importance in high-security applications where robust defenses are crucial, such as monitoring social media for harmful content or automating customer feedback analysis (Lavanya & Sosilaka, 2022; Piris & Gay, 2020), we focus on adversarial attacks against text-classification

Adversarial tactics range from simpler **character-level** modifications that introduce typographical errors which subtly confuse models but impair human readability. Similarly **word-level** changes replace individual words as seen in Table 1, to better deceive models while preserving semantic similarity but have limited word choices for effective perturbations. More complex **sentence-level** and **multi-level** approaches alter entire sentences or combine all three methods, greatly reducing classifier accuracy, but they are computationally intensive. This study analyzes word-level attacks for their balance between effectiveness, readability, and ease of implementation.

Sentence	Example	Prediction
Original	The **movie** was **awesome**.	Positive
Perturbed	The **film** was **amazing**.	Negative

Table 1: Word Level Text Classification Adversarial Attack Example

Recurrent Neural Networks (RNNs) and their variant, **Long Short-Term Memory Networks (LSTMs)**, once dominant in text classification, were limited by their sequential processing, which hindered the capture of long-range dependencies and increased computational demands (Fenglei et al., 2016). The introduction of Google's **Transformer** model in 2017, with its self-attention mechanism, revolutionized this by processing words simultaneously, enhancing dependency capture and parallelization efficiency (Vaswani et al., 2017). This advancement was further developed in the **Bidirectional Encoder Representations from Transformers (BERT)** model, which improved contextualization by interpreting words from both directions, thus deepening textual understanding (Devlin et al., 2019). Later models like Google's **T5** and OpenAI's **GPT** built on these innovations; T5 streamlined text processing for various tasks with

its text-to-text approach (Raffel et al., 2020), and GPT-3, pre-trained on 175 billion parameters, pioneered advanced zero-shot learning techniques (Brown et al., 2020; Bubeck et al., 2023). This study analyzes BERT due to its prominence in adversarial attack research and its reproducibility in text classification experiments.

Finally, this study focuses on **black-box** methods, which assess security without detailed internal model knowledge, underscoring their relevance in real-world applications. Such methods often predict success in **white-box** scenarios, where attackers have comprehensive insights into the model's parameters. We specifically analyze three types of black-box attacks:

1. **Score-based:** Utilizes the model's output confidence scores to craft perturbations that substantially affect classification outcomes.
2. **Decision-based:** Employs the model's final predicted classes to generate inputs that induce misclassification.
3. **Hybrid:** Integrates score and decision-based approaches, enhancing efficiency and reducing the need for extensive queries.

2 Recent Progress

Due to the aforementioned reasons we narrow our scope down to accurately judge the most impactful recent progress in adversarial attack research against text classification models and choose to look at significant breakthroughs in **black-box word-based adversarial attacks on BERT for text classification**. Therefore for the rest of this section it is assumed that the analysis provided is done for results on BERT even if the study incorporated other models in its exploration.

2.1 Score-Based Attack

Paper: Query-Efficient and Scalable Black-Box Adversarial Attacks on Discrete Sequential Data via Bayesian Optimization

A Blockwise Bayesian Attack (BBA) framework that employs Bayesian optimization with an Automatic Relevance Determination (ARD) categorical kernel to identify important positions in text is implemented in this study along with a post-optimization algorithm to produce adversarial examples with smaller perturbations (Lee et al., 2022). Average results over IMDB, Yelp, MR, AG and EC datasets achieve over 90% attack success rate with just a 15% perturbation rate and a notably low query count of 154 demonstrating a significant approach in reducing the number of queries required to generate effective adversarial examples using Bayesian optimization.

2.2 Decision-Based Attack

Paper: Bridge the Gap Between CV and NLP! A Gradient-based Textual Adversarial Attack Framework

Textual Projected Gradient Descent (T-PGD) framework that adapts optimization-based adversarial attack methods from the computer vision (CV) to NLP, adds continuously optimized perturbations to the embedding layer which are then amplified during forward propagation. The perturbed latent representations are then decoded using a masked language model to generate potential adversarial samples (Yuan et al., 2023). Average results over SST-2, Amazon and IMDB datasets achieve over 95% attack success rate with

high semantic similarity score of 0.86, while even reducing grammatical errors by -0.12 compared to the original model. This paper is significant as it successfully bridges the methodological gap between CV and NLP, enhancing the generation of fluent and grammatically correct adversarial samples and highlighting the potential of interdisciplinary approaches in machine learning (ML).

2.3 Hybrid Attack

Paper: Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment

This study implements the TEXTFOOLER framework, which uses a hybrid approach across various text-classification tasks and datasets to substitute key words with semantically similar alternatives for crafting impactful adversarial examples that maintain human readability (Jin et al., 2020). Selected for experiment reproduction, this research explores whether this hybrid method enhances the effectiveness and computational efficiency of adversarial attacks on text classification. The framework employs two primary techniques for generating adversarial examples.

1. **Word Importance:** Identifies the key words in a sentence that most influence the model's prediction and measures the importance of each word by observing changes in word importance score when words are removed.

Original Sentence: The movie was incredibly thrilling and well-acted.

Original Prediction: Positive (Score: 95.00%)

Removed Word	Modified Sentence	Prediction	Score (%)
The	movie was incredibly thrilling and well-acted	Positive	95.000000
movie	The was incredibly thrilling and well-acted	Positive	92.000000
was	The movie incredibly thrilling and well-acted	Neutral	89.000000
incredibly	The movie was thrilling and well-acted	Positive	80.000000
thrilling	The movie was incredibly and well-acted	Negative	60.000000
and	The movie was incredibly thrilling well-acted	Positive	91.000000
well-acted	The movie was incredibly thrilling and	Neutral	65.000000

Figure 1: Python snippet showing how different words impact score

2. **Word Replacement:** Replaces important words with semantically similar words that fit the context and maintain grammatical correctness

The objective of this experiment was to generate adversarial examples that modify BERT's predictions while preserving the semantic similarity to the original text across five datasets: AG, Fake, MR, IMDB, and Yelp. The successful reproduction of the results, as indicated in Table **??**, mirror the findings of the original study. Notably, the experiment achieved an 80% attack success rate coupled with a 14% perturbation rate and a semantic similarity score of 0.72, despite a high query count of 1377. These results highlight BERT's susceptibility to hybrid attacks, which efficiently produce high-quality adversarial examples with minimal perturbations. This shows the critical need for robust development practices in models deployed within sensitive applications, emphasizing the importance of enhancing their resilience against such sophisticated attacks.

Metrics	MR	IMDB	Yelp	AG	Fake	Average
Original Accuracy	86.00	90.90	97.00	94.20	97.80	93.18
After-Attack Accuracy	11.50	13.60	6.60	12.50	19.30	12.70
Attack Success Rate	74.50	77.30	90.40	81.70	78.50	80.48
% Perturbed Words	16.70	6.10	13.90	22.00	11.70	14.08
Semantic Similarity	0.65	0.86	0.74	0.57	0.76	0.72
Query Number	166	1134	827	357	4403	1377
Average Text Length	20	215	152	43	885	263

Table 2: Final Results

3 Conclusion

3.1 Pros and Cons of Current Methods

The recent advancements in adversarial text generation have demonstrated significant strides in maintaining semantic and grammatical integrity while ensuring high efficiency and transferability across different models and datasets. Methods such as Bayesian Optimization and TEXTFOOLER are noted for their ability to achieve high success rates with minimal modifications and reduced queries, showcasing enhanced applicability across various contexts. However, these methods also face challenges such as high computational demands, which may limit their practical viability. The effectiveness of these approaches can vary considerably depending on the target model and dataset, requiring extensive tuning. Additionally, the complexity inherent in techniques like Bayesian Optimization poses challenges for scalability and broad application.

3.2 Opinions After Analysis

The innovative application of techniques borrowed from machine learning and computer vision, like continuous perturbations and Bayesian Optimization, underlines the potential for breakthroughs in adversarial attack methodologies on text-classification systems. These studies underline the critical vulnerabilities in leading models such as BERT, highlighting an urgent need for the development of robust defense mechanisms. Future research should aim at simplifying these advanced methods to enhance their practicality and efficiency without sacrificing their effectiveness.

3.3 Future Work

Looking ahead, it is imperative to explore strategies that could reduce the computational overhead associated with these adversarial methods, simplifying optimization processes to facilitate broader application. There is also a significant need to enhance the fluency and grammatical correctness of the generated adversarial text, particularly for shorter texts. Developing more efficient query strategies that optimize the balance between exploration and exploitation could improve the feasibility of these methods in real-world scenarios. Additionally, integrating layered semantic checks could ensure that the generated adversarial text maintains a high level of semantic integrity.

4 References

1. Lavanya, P. M., & Sosilaka, E. J. (2022). Auto capture on drug text detection in social media through NLP from the heterogeneous data. Measurement: Sensors, 24, 100550. https://doi.org/10.1016/j.measen.2022.100550

2. Piris, Y., & Gay, A.-C. (2020). Customer satisfaction and natural language processing (NLP). Journal of Business Research, 123, 456-789. https://doi.org/10.1016/j.jbusres.2020.09.034

3. Fenglei, L., Xipeng, Q., & Xuanjing, H. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101. Retrieved from https://arxiv.org/abs/1605.05101

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Minneapolis, MN: Association for Computational Linguistics.

6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1-67. https://jmlr.org/papers/v21/20-074.html

7. OpenAI. (2023). GPT-4 Technical Report. Retrieved from https://openai.com/research/gpt-4

8. Bubeck, S., et al. (2023). Is GPT-4 a Good Data Analyst? arXiv preprint arXiv:2305.15038.

9. Lee, D., Moon, S., Lee, J., & Song, H. O. (2022). Query-Efficient and Scalable Black-Box Adversarial Attacks on Discrete Sequential Data via Bayesian Optimization. Proceedings of the 39th International Conference on Machine Learning, PMLR 162, 2022.

10. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20).

11. Yuan, L., Zhang, Y., Chen, Y., & Wei, W. (2023). Bridge the Gap Between CV and NLP! A Gradient-based Textual Adversarial Attack Framework. arXiv preprint arXiv:2110.15317v4.

12. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

13. Varghese, B., & Buyya, R. (2018). Next Generation Cloud Computing: New Trends and Research Directions. Future Generation Computer Systems, 79(Part 1), 849-861. https://doi.org/10.1016/j.future.2017.09.020