Investigation of Language Model Vulnerability Toward Adversarial Attacks and Exploration of Defense Techniques

Doeun Lee, Minjae Bae

Abstract

With the advancement of language models in sentiment analysis, ensuring their robustness against adversarial examples becomes essential to widen the scope of application. This project examined the susceptibility of three Transformer-based models – BERT, RoBERTa, and ELECTRA – to adversarial examples generated by TextFooler method, and significant deterioration in accuracy was observed for attacked models. To enhance the robustness, we implemented augmented training of original and adversarial examples, resulting in a reliable accuracy against both original and adversarial testing. To reflect the real-world problem of not having previous knowledge of adversarial data, we devised weighted voting of three original models. This achieved 76% accuracy while individual models' performance was around 64-70%. We expect these defense techniques to contribute to strengthening language models against real-world adversarial threat.

1 Introduction

With the advancement of language technologies, a growing number of users utilize language models for various tasks. Pre-trained models are particularly popular due to their applicability to general domains and the lack of necessity for time-consuming training. However, these systems face significant challenges in terms of robustness, particularly when exposed to adversarial examples (Szegedy et al., 2014).

Adversarial attacks have been shown to exploit vulnerabilities by introducing small, often imperceptible, perturbations that can lead to drastic misinterpretations of the input (Jia & Liang, 2017). These attacks pose a threat to the robustness of language models, undermining their trustworthiness in real-world applications. Investigating the vulnerability of models toward adversarial data could not only detect the weaknesses of the models but also propose evaluation metrics for future improvements.

The primary goal of this project is to analyze the impact of these attacks on three language models trained on a common dataset and explore the effectiveness of proposed defense mechanisms. We investigate two methods to mitigate the vulnerability and preserve the adaptability of the models:

- **Augmented training**
- **Ensemble models through weighted voting**

This report is organized as follows: Section 2 reviews related work on adversarial attacks and defense mechanisms for language models. Section 3 outlines the methodologies, including adversarial example generation and defense strategy implementation. Section 4 presents the evaluation results, demonstrating the effectiveness of the proposed defenses. Section 5 concludes with key findings and future research directions. Finally, Section 6 details the contributions of each team member.

2 Related Work

2.1 Adversarial Attacks in NLP

Adversarial attacks in NLP aim to subtly alter input texts to deceive language models without significantly changing the original meaning (Jia & Liang, 2017). Early work by Jia and Liang (2017) demonstrated that replacing key words with synonyms or semantically similar alternatives could mislead models into making incorrect predictions. Ebrahimi et al. (2018) introduced HotFlip, a white-box adversarial attack method that identifies critical characters in text for manipulation, further highlighting the susceptibility of NLP models to such attacks.

TextFooler, developed by Jin et al. (2020), is one of word-level attack tool for generating adversarial text by systematically substituting words to degrade model performance while maintaining grammatical correctness and semantic coherence.

2.2 Defense Mechanisms

Mitigating the vulnerability of language models to adversarial attacks is a critical area of research. Several defense strategies have been proposed to enhance the robustness of these models. Two prominent approaches are adversarial training and ensemble modeling.

Adversarial training is one of the most widely recognized methods for defending against adversarial attacks. This approach involves training models on a mixture of original and adversarial examples, thereby exposing the model to potential perturbations during the training process (Tramèr et al., 2020). By learning from both clean and adversarial data, the model can develop more robust representations that are less susceptible to deceptive inputs. Previous studies have demonstrated that adversarial training can significantly enhance a model's resilience against known adversarial examples (Liu et al., 2019). However, this method has limitations, as it primarily improves robustness against the specific types of adversarial data encountered during training, potentially leaving models vulnerable to novel or unforeseen attack vectors.

Ensemble modeling is another effective defense strategy that involves combining the predictions of multiple models to improve overall performance and robustness. Tramèr et al. (2020) proposed an ensemble adversarial training method that trains multiple models on different subsets of adversarial examples. This approach capitalizes on the diverse decision boundaries of individual models, making it harder for adversarial attacks to succeed across the entire ensemble. The ensemble's final prediction is typically determined through voting mechanisms, where each model contributes to the final decision based on its confidence or accuracy. This ensemble strategy not only enhances robustness against known adversarial attacks but also provides a safeguard against novel perturbations that may not have been explicitly encountered during training. The inherent diversity within the ensemble ensures that adversarial examples exploiting specific vulnerabilities in one model are less likely to deceive others, thereby maintaining overall classification accuracy.

2.3 Summary

The existing research highlights the significant challenges posed by adversarial attacks to the robustness of language models. Adversarial training offers a direct method to enhance model resilience by incorporating adversarial examples into the training process, while ensemble modeling provides a complementary approach by leveraging the strengths of multiple models to mitigate the impact of both known and unforeseen attacks.

Building on these foundational defense mechanisms, our project aims to analyze the susceptibility of BERT, RoBERTa, and ELECTRA to adversarial attacks and evaluate the effectiveness of augmented training and weighted voting ensemble methods in enhancing model robustness. By integrating insights from adversarial training and ensemble modeling, we seek to develop a comprehensive defense strategy that not only addresses known vulnerabilities but also anticipates and mitigates potential future adversarial threats.

3 Our Method

3.1 Adversarial Data Investigation

To assess the vulnerability of the selected language models, we employed TextAttack with TextFooler recipe (Jin et al., 2020), a comprehensive framework for generating adversarial examples in NLP tasks (Morris et al., 2020). TextFooler generates adversarial examples by identifying and substituting critical words in sentences with their synonyms or contextually similar alternatives, thereby altering the model's prediction without significantly changing the sentence's meaning.

For instance, to illustrate the impact of adversarial attacks on sentiment classification, consider the following example. In our experiment, as shown in **Figure 1**, the original sentence "much

better than I **expected**" was transformed into "much **greatest** than I **awaits**," resulting in a shift of the sentiment classification from positive to negative. This alteration exposed the model's sensitivity to specific word choices and highlighted its vulnerability to adversarial perturbations. By generating a substantial set of adversarial examples, we systematically evaluated the models' robustness and identified common patterns of vulnerability across different architectures.

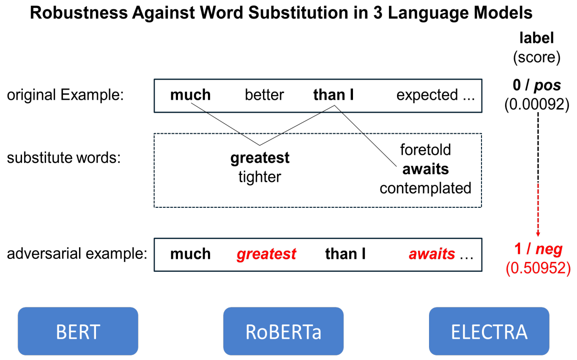

Figure 1: Example of Adversarial Attack on Sentiment Analysis. The original sentence "much better than I expected" is classified as **Positive**. After applying TextFooler, the sentence is modified to "much greatest than I awaits," resulting in a **Negative** classification.

Additionally, we generated three adversarial datasets, each by attacking a different model—BERT, RoBERTa, and ELECTRA—using TextFooler. Each adversarial dataset was then evaluated across all three models to comprehensively assess their robustness against targeted attacks. This approach allowed us to understand not only how each model responds to attacks specifically designed against it but also how resilient it remains when faced with adversarial examples generated for other models.

The evaluation results are summarized in **Table 1**, which presents the performance of each model on the adversarial datasets generated from the other models. The table highlights the cross-model vulnerability, indicating how an adversarial attack on one model can affect the performance of other models.

		Tested Model					
		BERT		RoBERTa		ELECTRA	
		ACC	F1	ACC	F1	ACC	F1
Attacked	Original	**89.01**	**89.13**	91.44	91.53	90.80	90.99
	BERT	**47.96**	**48.52**	85.36	86.07	82.72	83.64
	RoBERTa	68.40	66.81	**39.88**	**40.00**	75.76	75.78
	ELECTRA	75.44	73.67	84.92	84.38	**38.96**	**38.91**

Table 1: Performance of Models on Adversarial Datasets Generated from Different Attacked Models.

3.2 Augmented Training

Augmented Training is a defense mechanism aimed at enhancing model resilience by incorporating adversarial examples into the training process. As shown in **Figure 2**, we combined 60% of the original dataset with 40% adversarial data generated by TextFooler. BERT was fine-tuned on this mixed dataset to improve their ability to correctly classify both original and adversarial inputs.

This approach leverages exposure to adversarial perturbations during training, enabling the models to learn more robust representations and reduce the likelihood of misclassification when encountering similar adversarial attacks in deployment. The effectiveness of augmented training was evaluated by comparing the model's performance on both original and adversarial test sets before and after the training process.

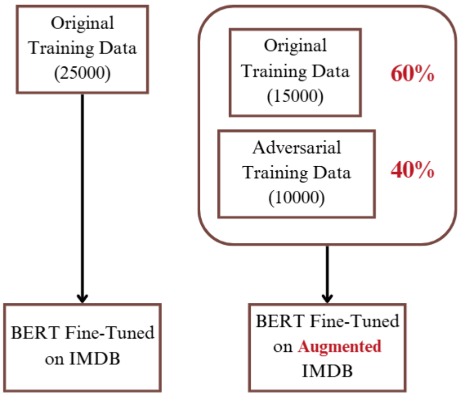
Original Training Data (25000)
Original Training Data (15000)
60%
Adversarial Training Data (10000)
40%
BERT Fine-Tuned on IMDB
BERT Fine-Tuned on Augmented IMDB

Figure 2: Augmented Training Process. The original dataset (60%) is combined with adversarial data (40%) to create a mixed training set. BERT is fine-tuned on this combined dataset to enhance its robustness against adversarial attacks.

The experimental results, summarized in **Table 2**, demonstrate that fine-tuning with augmented data leads to a significant increase in accuracy on adversarial data while only resulting in a minimal decrease in accuracy on the original dataset.

	Tested Model			
	BERT		BERT Augmented	
	ACC	F1	ACC	F1
Original	**89.01**	**89.13**	88.27	88.27
Adversarial	47.96	48.52	**83.80**	**84.24**

Table 2: Performance of BERT on Original and Adversarial Datasets

3.3 Ensemble Models Through Weighted Voting

The second defense strategy involves ensemble modeling through weighted voting. This method combines the predictions of multiple models—BERT, RoBERTa, and ELECTRA—by assigning weights to each model based on their individual performance. The final prediction is determined by the weighted sum of the individual model predictions, thereby leveraging the strengths of each model to achieve higher overall accuracy.

In real-world scenarios, it is often impractical to fine-tune models with augmented data or to determine which specific model is best suited for a given example. To address this, we propose using weighted voting of models fine-tuned solely with original data. This approach simplifies the deployment process by eliminating the need for additional fine-tuning based on augmented datasets and allows for a more generalized defense against diverse adversarial attacks.

We implemented a **weighted voting ensemble** approach to combine the predictions of BERT, RoBERTa, and ELECTRA. The ensemble assigns weights to each model based on their individual performance, allowing the final prediction to be determined by the ensemble of the models' predictions. Output logits of each data were used for weighted sum to reflect the probabilistic likelihood of each classification outcome rather than relying on binary class labels with fixed values. This method helps mitigate the weaknesses of individual models and enhances robustness against adversarial attacks. The final classification is determined as class with the highest weighted sum of logits.

Figure 3 illustrates the pseudo-code for the weighted voting ensemble method used in our experiments.

```
Assign Weights Based on Accuracies;
    $w_m \leftarrow \frac{\mathrm{Acc}_m^{\mathrm{train}}}{\sum_{m' \in Models} \mathrm{Acc}_{m'}^{\mathrm{train}}} \quad \forall m \in Models;$
Logit Extraction;
    $\mathcal{L}_m^{\mathrm{test}} \leftarrow \mathrm{GetLogits}(\mathrm{Model}, \mathrm{Data});$
Weighted Logit Calculation;
    $\mathcal{L}_{\mathrm{weighted}}^{\mathrm{test}} \leftarrow \sum_{m \in Model} w_m \cdot \mathcal{L}_m^{\mathrm{test}};$
Final Weighted Predictions;
    $\hat{y}^{\mathrm{weighted}} \leftarrow \arg\max \left( \mathcal{L}_{\mathrm{weighted}}^{\mathrm{test}} \right);$
```

Figure 3: Pseudo-Code for Weighted Voting Ensemble.

This ensemble method was tested using the adversarial data from all three adversarial datasets generated during the adversarial data exploration phase. The results demonstrated improved performance on adversarial data compared to individual models while maintaining high accuracy on original data.

4 Evaluation

4.1 Results

The weighted voting ensemble achieved the highest overall accuracy of 0.76, outperforming all individual models—BERT (0.64), RoBERTa (0.66), and ELECTRA (0.70). As shown in **Figure 4**, the ensemble model also achieved balanced performance across positive and negative F1-scores, demonstrating its robustness against adversarial attacks. This outcome highlights the effectiveness of combining multiple models to mitigate vulnerabilities inherent in individual models.

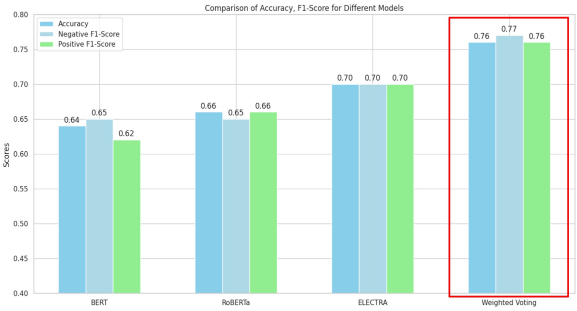

Figure 4: Comparison of Accuracy and F1-Scores for BERT, RoBERTa, ELECTRA, and the Weighted Voting Ensemble.

4.2 Impact of Weighted Voting Ensemble

The weighted voting ensemble achieved the highest accuracy of 0.76, surpassing all individual models. This outcome underscores the advantage of combining multiple models to leverage their diverse strengths, resulting in a more robust and accurate classification system. The ensemble model proved particularly effective in mitigating the impact of adversarial attacks, as it reduced the likelihood of a single model's vulnerability affecting the overall prediction.

4.3 Analysis

The evaluation results confirm the effectiveness of the proposed defense mechanisms. **Augmented**

training, applied to BERT, significantly improved its robustness against adversarial attacks. While the augmented model showed a slight decrease in performance on the original dataset, it exhibited a significant improvement on adversarial data. Specifically, the accuracy increased from 47.96% to 83.80%, and the F1-score improved from 48.52 to 84.24, highlighting the augmented model's enhanced robustness against adversarial attacks.

One key advantage of the **weighted voting** ensemble is its ability to operate effectively without prior knowledge of which model was targeted during the creation of adversarial examples. This property makes the ensemble approach particularly valuable in real-world scenarios, where adversarial data is often not the result of deliberate attacks but rather naturally generated (e.g., by human error, language variations, or informal communication). In such cases, determining which model to exclude or prioritize is infeasible, making weighted voting a dependable and practical metric for sentiment classification.

This adaptability underscores the ensemble's robustness as a realistic solution for handling unstructured and unpredictable adversarial data in practical applications. By relying on the strengths of multiple models and balancing their predictions, the ensemble method minimizes the impact of individual model vulnerabilities and ensures consistent performance across diverse datasets.

5 Conclusions

This project investigated the vulnerability of Transformer-based language models—BERT, RoBERTa, and ELECTRA—to adversarial attacks and explored two defense techniques: **augmented training** and **weighted voting** ensemble modeling. Through the generation of adversarial examples using TextAttack and subsequent evaluation, we demonstrated that both defense methods significantly enhance model robustness.

Augmented training improved the robustness of individual models against adversarial inputs, while the weighted voting ensemble provided a practical and adaptable solution for real-world scenarios. The ensemble approach demonstrated its effectiveness in handling adversarial data without requiring prior knowledge of which model was targeted, making it a reliable defense mechanism for naturally occurring adversarial inputs.

Future research will focus on a deeper analysis of the specific perturbations introduced in adversarial data to better understand the unique vulnerabilities of the three Transformer-based models investigated in this project. By examining the generated adversarial examples, we aim to identify the perturbations that were most effective against each model, providing valuable insights into their weaknesses. This analysis will guide the development of more targeted defense mechanisms to enhance model robustness in diverse and dynamic environments. Additionally, exploring advanced ensemble techniques and defense strategies, such as defensive distillation and input preprocessing, could further improve the resilience of language models, making them more reliable for real-world applications.

6 Team member's work and contributions

- Doeun
    - ■ Adversarial example generation and testing (BERT, RoBERTa)
    - ■ Weighted voting
- Minjae
    - ■ Adversarial example generation and testing (ELECTRA)
    - ■ Augmented training
- Equal contribution or done together
    - ■ Report writing and presentation preparation
    - ■ Project management and coordination
    - ■ Literature review and related work compilation
    - ■ Performance evaluation and analysis

References

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. *Intriguing properties of neural networks.* In arXiv preprint arXiv:1312.6199.

Robin Jia and Percy Liang. 2017. *Adversarial examples for evaluating reading comprehension systems.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark, Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. *HotFlip: White-Box Adversarial Examples for Text Classification.* In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), edited by Iryna Gurevych and Yuji Matsumoto, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. *Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment.* In arXiv preprint arXiv:1907.11932.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2020. *Ensemble Adversarial Training: Attacks and Defenses.* In arXiv preprint arXiv:1705.07204.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. *Inoculation by fine-tuning: A method for analyzing challenge datasets.* 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), pages 2171–2179, Minneapolis, Minnesota, June. Association for Computational Linguistics.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).