Generating and Comparing Adversarial Attacks on text based sentiment detector using Integrated Gradient

Pulkit Tyagi1, Manish Patro2, Vikash Kumar Arya3, Akash Bera4, Dr. Chittaranjan Swain5

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

imanish0409@gmail.com

hello.kvikash@gmail.com

tyaggipulkit@gmail.com

2041002091.akashbera@gmail.com

**Abstract**

This project report details our exploration of adversarial robustness in text-based sentiment analysis. We investigated the vulnerabilities of these models to adversarial manipulations and evaluated various attack techniques to enhance the robustness of our sentiment detector. Our primary focus was on leveraging Integrated Gradients (IG) as a novel approach to generate more effective defenses against adversarial perturbations. Through extensive experimentation, we compared the performance of our sentiment analysis model trained on datasets perturbed by various attack algorithms, including Fast Gradient Sign Method (FGSM), K-Projected Gradient Descent (K-PGD), and Integrated Gradients. We analyzed the impact of these perturbations on the model's accuracy, loss, hits, misses, elapsed time, and precision. The results highlight the significant impact of adversarial training on model robustness and shed light on the advantages and limitations of different attack techniques. We found that while K-PGD demonstrated superior accuracy and loss minimization, IG emerged as a more efficient and scalable approach, offering a promising alternative for enhancing model robustness while maintaining computational efficiency. This report serves as a comprehensive document outlining our research methodology, experimental findings, and conclusions. It provides valuable insights into the challenges and opportunities presented by adversarial attacks in text-based sentiment analysis, contributing to the ongoing efforts to develop more reliable and trustworthy AI systems.

**Keywords:**  Natural Language Processing, Machine Learning

1. Introduction

Natural Language Processing (NLP) has made many significant advancements in the recent years with neural network models making major feats in performing brilliantly on a variety of tasks. This is an area of technology that is in a constant state of change with developments happening nearly by the minute. Sentiment analysis, that is the task of identifying and classifying the emotional tone expressed in text (for example, “this movie is not of my liking”, the tone expressed here is that of being disappointed or in other words a negative emotion), has become increasingly important in various domains, including customer feedback analysis, social media

monitoring, and opinion mining.

However, despite these advancements a major weakness of these models is how easily these models can be manipulated by making subtle changes in the input that leads to drastic change in the output, compromising the predictions made by the model. The susceptibility of all of these models to adversarial attacks reduces the robustness and reliability of these models. Adversarial attacks take advantage of this key weakness in neural networks and make slight, undetectable but acutely formed modifications to text that they feed to models and this can cause the models to output wrong and inevitably incorrect outputs and lead them to make compromised decisions. This vulnerability is a major problem for NLP systems that need to be deployed in the real world as part of applications where trust and correctness is a major concern.

Consider, for instance, a scenario where a sentiment analysis model is deployed to analyze customer reviews for a product. An attacker could manipulate the reviews by subtly altering the wording, effectively "poisoning" the dataset and leading the model to misinterpret the sentiment, potentially impacting product reputation and sales.

This project delves into the critical issue of adversarial robustness in text-based sentiment analysis models. Our work focuses on comparing and evaluating various attack techniques across multiple criteria in order to strengthen the text-based sentiment analysis model. We aim to investigate the vulnerabilities of these models to

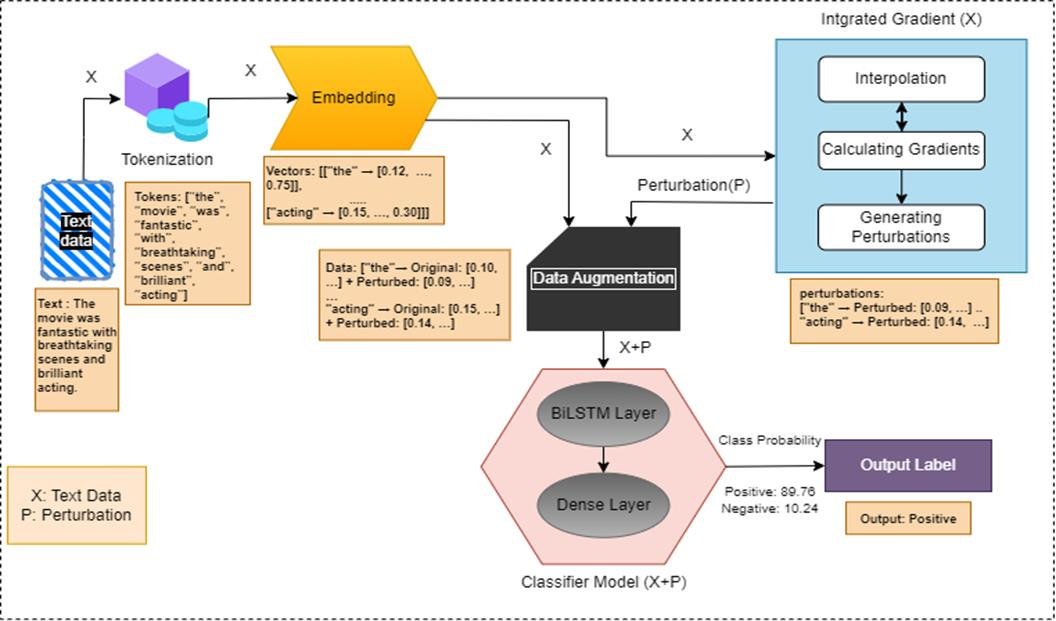
adversarial manipulations and to discover practical strategies for fortifying their defenses against such assaults. Our primary objective is to contribute to the development of more robust and trustworthy NLP systems that can withstand adversarial manipulations and maintain their accuracy and reliability in real-world applications.

1. Literature Survey

In the realm of adversarial attack algorithms, FGSM, PGD, and IG have been very instrumental. FGSM, introduced in 2014[17], is the simplest and most straightforward to be implemented. It forms one of the fundamental methods that produce weak and less stable perturbations. PGD, introduced in 2017[7], made a large improvement over FGSM by using gradient iterative updates with projection, often the case when looking for both robust and effective adversarial examples, but at higher computational costs. IG, introduced in 2017[8], has a path-based approach to integrating gradients, ensuring more comprehensive and stable perturbations but offering a much more interpretable approach compared to FGSM and PGD. It can thus balance the strength and computational cost quite well, which suffers a lot from the overfitting problem in PGD[18]. Indeed, studies claim that IG enables the construction of strong and stable adversarial examples with no overfitting and is thus, very promising for improving model robustness and understanding vulnerabilities. In most recent studies, performance comparison has been made on metrics such as accuracy and attack success rate. However, even with advances, the need remains to have techniques which create robust and generalizable adversarial examples without fitting; this underlines the need for continuing the research in methods such as IG.

1. Proposed Model

For this work we compared three attack algorithms, IG, FGSM, K-PGD to get an insight on how the model’s accuracy and robustness is affected when the model is trained on the perturbed dataset generated by the mentioned attack algorithms. IG is implemented to generate perturbations for the embedded vector.



**Fig 1: BiLSTM Sentiment Analyzer with IG-Enhanced Robustness**

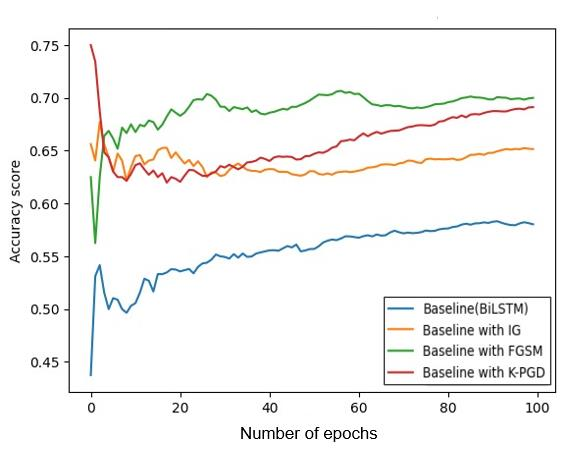
Firstly, the text is converted into a sequence of numerical indices, representing individual words or sub words using Text-Vectorization. The numerical index transformed into a dense vector (embedding vector), effectively turning words into meaningful numerical representations that capture semantic relationships in embedding layer.

The Integrated Gradients is implemented to generate interpolation of embeddings moderately moving from a baseline embedding (filled with zeroes) to the target embeddings. The gradients are calculated using these interpolated embeddings and we calculated component wise mean of that gradient for the model’s output probabilities. We then integrate those gradients using Reimann summation [8] to generate perturbation. The input embeddings and the perturbation are combined in the augmented layer for the training of the model. Finally, the model is trained through the perturbed embeddings and generate probabilities for each class. The model consists of Bidirectional Long Short-Term Memory (BiLSTM) with 64 units (64 is the more flexible and efficient for representing the data) and Dense layers.

1. Experimentation and Model Evaluation

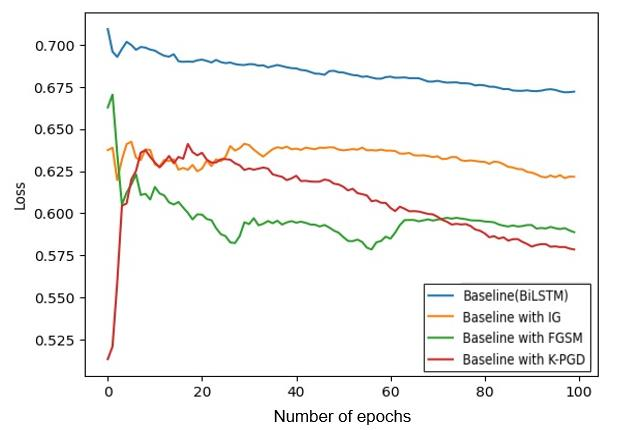
Our evaluation measures focused on assessing the model's performance in both standard classification accuracy and its robustness against adversarial attacks. We compared the different algorithm for everymetrics (i.e. IG, KPGD, FGSM and baseline BiLSTM) with each other for evaluations.

For result analysis we are comparing four models, a baseline model which is a BiLSTM model, baseline model which is trained on the perturbed dataset created using FGSM, baseline model which is trained on the perturbed dataset created using PGD and baseline modal which is trained on the perturbed dataset using IG. All the models are trained for the same number of epochs and compared using same metrics.



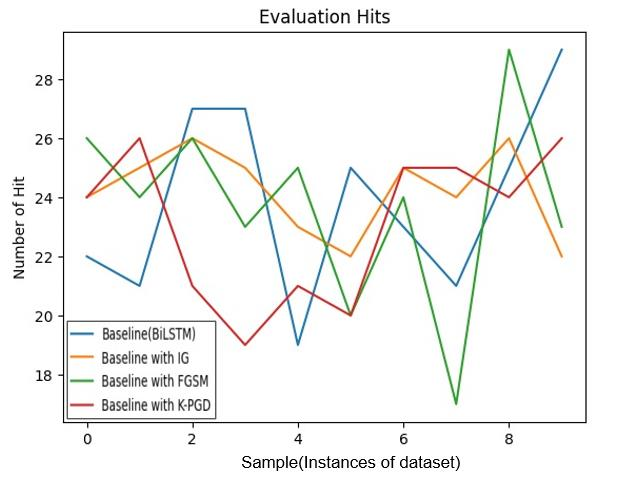
**Fig 2: Accuracy score comparison between different algorithms**

The graph in Figure 2 here clearly represents that the adversarial training using K-step Projected Gradient Descent (K-PGD) significantly improves the model’s accuracy, hence making it the most effective method for enhancing robustness. IG training on the other hand showcases a similar accuracy to the baseline model, which indicates that IG successfully generated improved adversarial examples that do not hinder the learning process. FGSM showed inconsistent and less effective trends, simply meaning it appears less effective in improving robustness, as its accuracy fluctuates more, suggesting that it has less robust adversarial examples. Baseline model showed the least accuracy, about 0.57, whereas baseline model that was trained on K-PGD showed the highest accuracy, about 0.69.



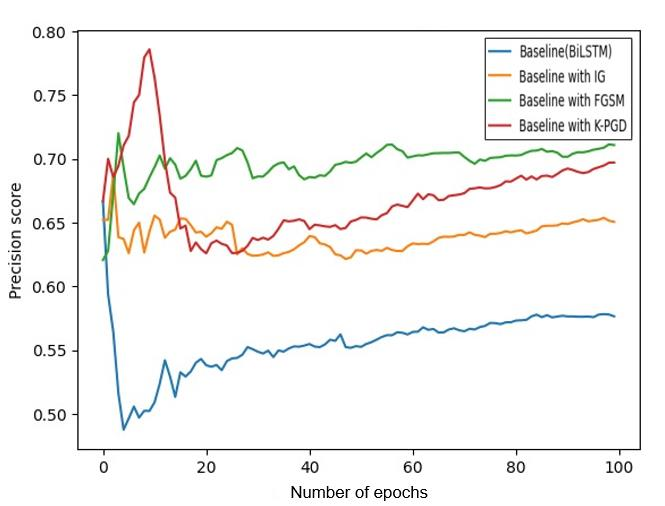
**Figure 3: Loss score comparison between different algorithms**

K-PGD appears to be the most effective adversarial training method in this scenario as its loss curve shows that it descends consistently and reaches the lowest value. IG seems to be a good balance between effectiveness and learning ability, its loss curve shows a positive descent that signifies that the perturbations created were effective and hence it doesn’t disrupt the training process. The inconsistency in the curve of FGSM indicates that the adversarial examples generated by the algorithm were weak and confusing to the model. Lowest loss was showed by K-PGD and the highest loss value was showed by the baseline model.



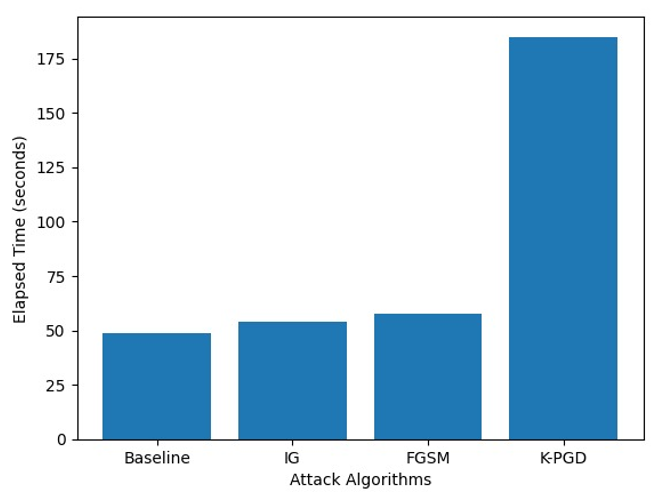
**Figure 4: Comparison of Hit metric amongst different algorithms**

The inference that we get out of this graph shown in Figure 3 can be summarised as, K-PGD performing the best over this metric, followed by IG, FGSM and lastly the baseline modal. The Hit metric is nothing but the number of times the model correctly predicted the sentiment of the adversarial example. IG demonstrates a generally higher number of hits, indicating that it has improved its robustness. FGSM shows a more erratic performance, with higher number of hits for some samples but a lower number for others, indicating that FGSM training might have a mixed effect on robustness which indirectly means that adversarial perturbations created by FGSM were weak and inconsistent. K-PGD demonstrates the highest number of hits across most of the samples signifying that it has significantly enhanced the model’s robustness against adversarial examples.



**Figure 5: Precision being used as a metric for comparison between algorithms**

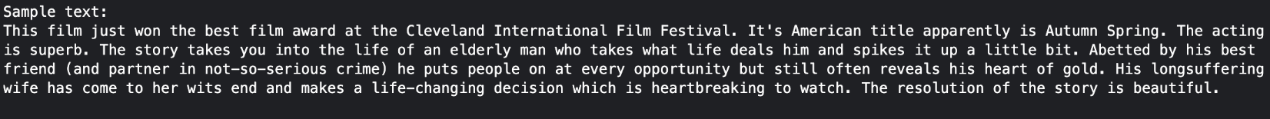
Based on the graph, FGSM appears to be marginally better than IG in terms of achieving a higher precision score. However, IG provides a more stable precision over the training iterations. The K-PGD method, despite initial fluctuations, ultimately achieves the highest precision score, suggesting it is the most effective in enhancing the model's robustness and precision.



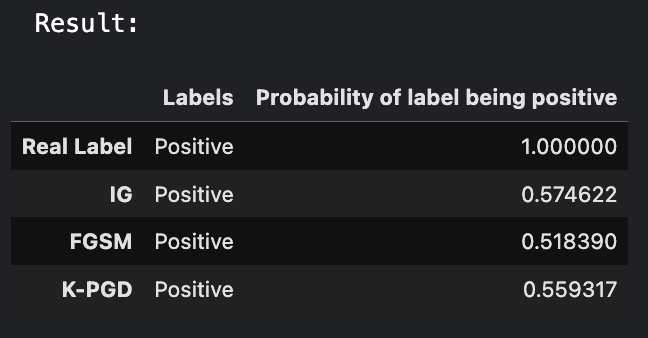
**Figure 6: Comparison of Elapsed Time amongst different algorithms**

The Elapsed Time taken by K-PGD is quite high when compared to that of IG and FGSM. The graph shows a trade-off between computational efficiency and the robustness provided by the adversarial training techniques. While K-PGD may provide better robustness (as seen in accuracy scores), it comes at a higher computational cost.The higher elapsed time for K-PGD indicates that it is computationally more intensive, likely due to its iterative nature, where multiple steps are taken to refine the adversarial examples.

We also performed a evaluation of Integrated Gradients by testing sentiment of a sample text by generating different adversaries from the algorithms.



**Figure 7: Sample text**



**Figure 8: Result of label probabilities of being positive by every algorithm**

In comparison to other algorithms the model trained with Integrated gradients adversaries are showing a bit better probability score than FGSM and K-PGD.

1. Conclusion and Future Scope

The main work of our project was to compare different adversarial training techniques namely IG, FGSM, KPGD, in enhancing the robustness of a text-based sentiment analysis model. The model that we used was BiLSTM model, and we trained the model on multiple perturbed datasets. Our findings demonstrate the impact of adversarial training on model performance, with K-PGD emerging as the most effective method in improving model accuracy, minimizing loss and maximizing the number of correct classifications on adversarial examples.

FGSM’s performance over multiple metrics was very inconsistent and the fluctuations in the trends potentially indicated that FGSM generated weaker perturbations which in turn provided less robust adversarial examples for the model to train on. While IG showed a more balance approach it was better than FGSM in most comparisons and was matching with K-PGD over almost all metrics. IG had an upper hand over both FGSM and K-PGD in the Elapsed Time metric showing that it was much less computationally expensive both in time and computational resources. K-PGD suffers from a drawback of overfitting, which in simple terms means that it won’t have consistent performance over different models and datasets. IG solves this problem with ease as it is a more scalable and generalizable algorithm as compared to the other two. Therefore, even though IG’s performance was slightly lesser than that of K-PGD’s, but due to its scalability to larger datasets and more

generalizability Integrated Gradients is a much better options for the generation of adversarial attacks and for increasing the robustness of NLP models.

The dataset that we used was not a large one therefore the results that came out were more leaned towards KPGD, but for a larger dataset and more a variety of models, IG can possibly be a better option for adversarial generation and model security.

References

[1] Li Dianqi, Zhang Yizhe, Peng Hao, Chen Liqun, Brockett Chris, Ting Sun Ming, Dolan Bill(2020).

"Contextualized perturbation for textual adversarial attack." Proceedings of the 2020 Conference on

Empirical Methods in Natural Language Processing (EMNLP).pp.8337–8347.

[2] Alzantot Moustafa, Sharma Yash, Elgohary Ahmed, Ho Bo-Jhang, Srivastava Mani B., Chang KaiWei(2018). "Generating Natural Language Adversarial Examples." Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2018).pp.2890–2900.

[3] X. Morris John, Lifland Eli, Yong Yoo Jin, Grigsby Jake, Jin Di, Qi Yanjun(2020). "TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).pp. 2889–2900.

[4] Wang Lina, Chen Xingshu, Tang Rui, Yue Yawei, Zhu Yi , Zeng Xuemei , Wang Wei(2020).

"Improving adversarial robustness of deep neural networks by using semantic information."

Proceedings of the AAAI Conference on Artificial Intelligence, 34.pp.5628–5635.

[5] Pan Lin, Hang Chung-Wei, Sil Avi, Potdar Saloni(2022). "Improved Text Classification via

Contrastive Adversarial Training." Proceedings of the AAAI Conference on Artificial Intelligence,

36.

[6] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair,

Sherjil; Courville, Aaron; Bengio, Yoshua (2014). "Generative Adversarial Nets". Proceedings of the International Conference on Neural Information Processing Systems (NIPS'14). pp 2672–2680.

[7] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). "Towards Deep Learning

Models Resistant to Adversarial Attacks." Proceedings of the International Conference on Learning

Representations (ICLR) , pp. 1-28.

[8] Sundararajan, M., Taly, A., & Yan, Q. (2017). "Axiomatic Attribution for Deep Networks."

Proceedings of the 31st International Conference on Neural Information Processing Systems

(NeurIPS), pp. 5575-5585.

[9] Wang, Y., Liu, J., Chang, X., Mišić, J., & Mišić, V. B. (2020). "IWA: Integrated Gradient based

White-box Attacks for Fooling Deep Neural Networks." Proceedings of the 2020 IEEE Conference

on Computer Vision and Pattern Recognition (CVPR), pp. 3678-3686.

[10] Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). "Defending Against Adversarial Attacks by Attributing Feature Importance." Proceedings of the 35th International Conference on Machine Learning (ICML), vol. 80, pp. 960-968.

[11] Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., & Jain, A. K. (2020). "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review." arXiv:1909.08072v3 [cs.LG], vol. 2020, pp. 1-35.

[12] Bhagoji, A. N., He, W., Li, B., & Song, D. (2018). "Exploring the Space of Black-box Attacks on Deep Neural Networks." Proceedings of the 2018 ACM SIGSAC Conference on Computer and

Communications Security (CCS), pp. 1310-1325.

[13] Wallace, E., Feng, S., Kandpal, N., & Gardner, M. (2020). "Universal Adversarial Triggers for

Attacking and Analyzing NLP." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2153-2162.

[14] Li, J., Ji, S., Du, T., Li, B., & Wang, T. (2019). "TextBugger: Generating Adversarial Text Against Real-world Applications." Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS), pp. 1-15.

[15] Huang, T., Menkovski, V., Pei, Y., & Pechenizkiy, M. (2021). "Bridging the Performance Gap

between FGSM and PGD Adversarial Training." Proceedings of the 2021 IEEE International

Conference on Big Data (Big Data), pp. 1-10.

[16] Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., & Liu, J. (2020). "FREELB: Enhanced

Adversarial Training for Natural Language Understanding." Proceedings of the 2020 Conference on

Empirical Methods in Natural Language Processing (EMNLP), pp. 8337-8347.

[17] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). "Explaining and Harnessing Adversarial

Examples." https://arxiv.org/abs/1412.6572.

[18] Tian, R., & Mao, Y. (2022). "On Robust Overfitting: Adversarial Training Induced Distribution

Matters." Proceedings of the 2022 International Conference on Learning Representations (ICLR),

pp. 1-12.