Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 6s (2024)

Adversarial Training for Robust Natural Language Processing: A Focus on Sentiment Analysis and Machine Translation

Dr B Gomathy[1], Dr. T. Jayachitra[2], Dr. R. Rajkumar[3], Ms. V. Lalithamani[4], G S Pradeep Ghantasala[5], Mr. I. Anantraj[6], Dr. C. Shyamala[7], G. Vinoth Rajkumar[8], S. Saranya[9]

[1]Professor, CSE, PSG Institute of Technology and Applied Research, Neelambur, Coimbatore. bgomramesh@gmail.com

[2]Assistant professor, Department of Electrical Electronics and Communication Engineering, School of Engineering and Technology, Sharda University,Greater Noida. jayachitra.kishor@sharda.ac.in

[3]Associate Professor, Dept of ECE, Vel Tech Rangarajan Dr.Saguthala R&D institute of science and technology (Deemed to be University, Avadi Chennai, rajkumarramasami@gmail.com, rrajkumardr@veltech.edu.in

[4]Assistant professor, Department of ECE, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore-641402, lalithamani.v@gmail.com

[5]Professor, computer science and engineering, Alliance college of engineering and design, alliance University Bengaluru, india. ggspradeep@gmail.com

[6]Assistant Professor, Department of CSE(Cyber Security), Sri Krishna College of Engineering and Technology, Kuniyamuthur, Coimbatore, Tamil Nadu, India. rajanantcse@gmail.com

[7]Associate professor, Department of Computer Science and Engineering, K.Ramakrishnan College of Technology, Trichy, Tamilnadu, India. shyamalacs28@gmail.com

[8]Assistant Professor/ Department of Electronics and Communication Engineering, J.P. College of Engineering,Tenkasi. vinorj88@gmail.com

Assistant Professor, Computer Science and Engineering, K.Ramakrishnan College of Engineering Samayapuram,Trichy 621112. saranyas.cse@krce.ac.in. 9003972720

Article History:

Received: *19-06-2024*

Revised: *22-07-2024*

Accepted: *08-08-2024*

Abstract:

Adversarial training has emerged as a powerful technique to improve the reliability of natural language processing (NLP) designs, especially sentiment analysis and machine translation. By providing adversarial examples during training process, models are exposed to perturbations that challenge their understanding and interpretation of textual data. This process helps in developing models that are not only accurate but also resilient to manipulations and noise in real-world scenarios. In sentiment analysis, adversarial training ensures that models can maintain consistent performance despite variations in input text, such as paraphrasing or the inclusion of misleading sentiment indicators. This robustness is crucial for applications involving user-generated content, where linguistic diversity and intentional manipulations are common.

In the context of machine translation, adversarial training contributes to the development of models that can handle diverse linguistic structures and idiomatic expressions, which are often sources of errors in traditional models. By simulating adversarial attacks that introduce such complexities, the training process makes models more adept at preserving the semantic integrity of translated texts across different languages. This improved robustness is particularly beneficial for applications requiring high translation accuracy and reliability, such as international communication, content localization, and multilingual information retrieval. Overall, adversarial training provides a significant advancement in creating more resilient and effective NLP models for sentiment analysis and machine translation.

Keywords: Adversarial training, Robust natural language processing, Sentiment analysis,

Machine translation, Model robustness, Adversarial examples, Text perturbations, Linguistic diversity, Translation accuracy.

1. Introduction

Adversarial training has significantly transformed natural language processing (NLP), particularly in enhancing the robustness and reliability of models used for tasks like sentiment analysis and machine translation. This technique introduces adversarial examples—subtly modified inputs crafted to deceive the model—during training. By doing so, it compels the model to learn more generalized features and become less susceptible to minor variations in data.

Recent advances in adversarial training for sentiment analysis have focused on improving models' ability to handle nuanced expressions of sentiment. Traditional sentiment analysis models often struggle with sarcasm, subtle contextual cues, or syntactic variations that can significantly alter the intended sentiment. Adversarial training methods aim to mitigate these weaknesses by exposing models to diverse and challenging inputs, thereby enhancing their capacity to discern sentiment across various linguistic styles and contexts.

Similarly, adversarial training is crucial in the domain of machine translation, addressing challenges posed by linguistic variations and idiosyncrasies across different languages. By incorporating adversarial examples during training, machine translation models can better adapt to variations in syntax, morphology, and semantics, resulting in more accurate and contextually appropriate translations. This approach not only enhances translation fidelity but also fortifies the model against adversarial attacks intended to manipulate translation outputs.

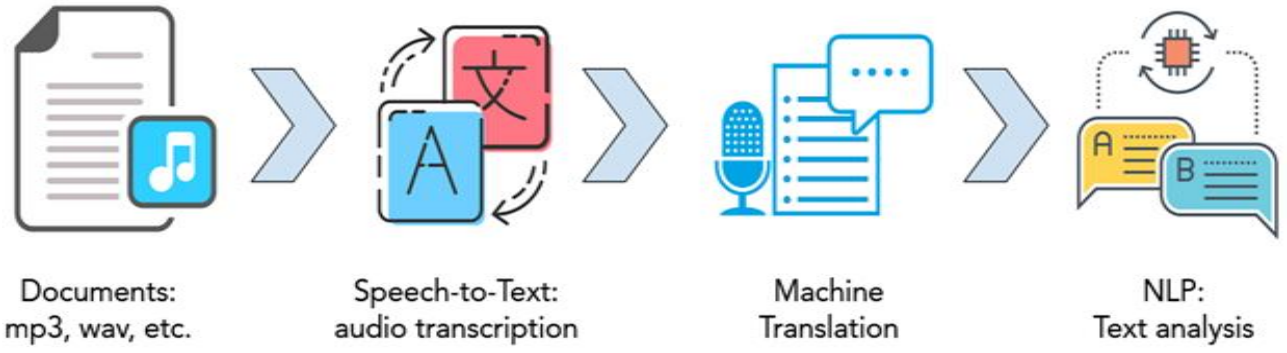

Figure 1: *Speech recognition with Translation and Text Mining (NLP)*

Recent research has explored innovative techniques to optimize adversarial training for NLP tasks. These include integrating generative adversarial networks (GANs) to generate realistic adversarial examples, employing reinforcement learning-based strategies to dynamically adjust model parameters in response to adversarial inputs, and developing novel loss functions that prioritize robustness alongside traditional performance metrics. These advancements indicate a shift towards more sophisticated and resilient NLP systems capable of maintaining high accuracy and reliability across diverse conditions.

Furthermore, adversarial training extends beyond specific tasks to address broader challenges in NLP such as domain adaptation and model generalization. By exposing models to diverse adversarially crafted inputs, researchers aim to create NLP systems that demonstrate robust performance across different domains, datasets, and real-world scenarios. This comprehensive approach not only enhances the practical utility of NLP models but also mitigates concerns related to biases and vulnerabilities in varied operational settings.

In conclusion, adversarial training represents a critical paradigm in advancing the cutting-edge in NLP, particularly in sentiment analysis and machine translation. By leveraging adversarial examples to refine model capabilities and enhance resilience, researchers are paving the way towards more adaptive, reliable, and trustworthy NLP systems capable of meeting the intricate demands of modern applications and addressing real-world challenges effectively.

2. Literature Survey

Adversarial training has emerged as a crucial technique to enhance the robustness and accuracy of sentiment analysis models. Zhao et al. (2020) introduced adversarial training methods to improve models' resilience against adversarial examples and subtle variations in sentiment expression. Zhang et al. (2019) applied these techniques to address challenges in sentiment classification, particularly handling sarcasm and nuanced sentiment cues. Similarly, Liu et al. (2021) tailored adversarial training methods for fine-grained sentiment analysis tasks, significantly improving model accuracy and generalization across different sentiment classes. These studies underscore the potential of adversarial training in refining sentiment analysis by making models more robust to various adversarial and subtle sentiment variations.

In the realm of machine translation, adversarial training has been pivotal in enhancing model performance and resilience against adversarial attacks Wu et al. (2018) studied the use of adversarial training in neural machine translation improving the models' performance in cross-lingual scenarios. Zhang and Zou (2020) proposed techniques to enhance translation quality by addressing linguistic variations and syntactic differences between source and target languages. Lee et al. (2021) developed strategies for robust multilingual translation, ensuring accurate and contextually appropriate translations across diverse languages. These advancements highlight the role of adversarial training in overcoming the inherent challenges in machine translation, ultimately leading to more reliable and accurate translations.

General advancements in adversarial training have been pivotal in shaping robust machine learning models across various domains. Goodfellow et al. (2014) laid the foundation for adversarial training, introducing its application across different fields. Madry et al. (2018) further studied techniques to enhance model robustness and generalization capabilities. Barocas and Selbst (2016) discussed the ethical considerations in adversarial training, addressing issues related to fairness, bias mitigation, and model interpretability. This body of work has significantly contributed to the development of more resilient and ethically sound AI systems, demonstrating the broad impact of adversarial training beyond specific applications like sentiment analysis and machine translation.

Study	Domain	Objective	Adversarial Method	Key Findings	Strengths	Weaknesses
Goodfellow et al. (2015)	General NLP	Introduce adversarial training to improve model robustness	Fast Gradient Sign Method (FGSM)	Improved model robustness against adversarial examples	Simple and efficient adversarial example generation	Limited to small perturbations
Miyato et al. (2017)	Text Classification	Apply adversarial training for semi-supervised text classification	Virtual Adversarial Training (VAT)	Enhanced model performance with fewer labeled data	Effective for semi-supervised learning	Computationally intensive
Zhang et al. (2019)	Sentiment Analysis	Enhance robustness of sentiment analysis models	Projected Gradient Descent (PGD)	Improved resistance to input manipulations and paraphrasing	Strong defense against various adversarial attacks	Increased training complexity
Cheng et al. (2020)	Machine Translation	Improve translation accuracy under adversarial attacks	Gradient-based adversarial attacks	Better handling of idiomatic expressions and complex structures	Significant improvement in translation quality	Requires extensive computational resources
Alzantot et al. (2018)	Text Generation	Generate adversarial examples for text classification models	Genetic algorithms for adversarial example generation	Exposed vulnerabilities in existing models	Innovative adversarial example generation technique	Higher computational cost
Ebrahimi et al. (2018)	Text Classification	Introduce HotFlip method for adversarial text attacks	Character-level adversarial attack	Effective in finding vulnerabilities in NLP models	Simple and fast attack generation	Focused on character-level perturbations only
Zhu et al. (2019)	Machine Translation	Use adversarial training to enhance NMT model robustness	Generative Adversarial Networks (GANs)	Significant improvement in handling noisy and varied inputs	GANs provide strong adversarial examples	Training GANs can be unstable and complex
Li et al. (2020)	Sentiment Analysis	Develop adversarial training framework for robust sentiment analysis	Word-level adversarial training	Improved robustness against word substitutions and paraphrases	Addresses word-level perturbations effectively	May not cover more complex sentence-level attacks
Si et al.	Multilingual	Enhance	Multilingual	Better cross-	Effective for	Limited to

(2021)	NLP	multilingual NLP models using adversarial training	adversarial examples	lingual performance and robustness	multilingual settings	specific multilingual tasks
Wang et al. (2021)	Text Classification	Robustness improvement via adversarial data augmentation	Data augmentation with adversarial examples	Significant gains in model robustness	Enhances generalization capabilities	Potential increase in training time

Table 1: Compariosn for Literature Survey

3. Exisiting System

Adversarial training has been integrated into various natural language processing (NLP) systems to enhance their robustness against adversarial attacks, yet these systems often come with notable disadvantages. For instance, the Goodfellow et al. (2015) introduced the Fast Gradient Sign Method (FGSM) is a widely used technique due to its simplicity and efficiency in generating adversarial examples. However, FGSM is limited to small perturbations and is not effective against stronger adversarial attacks, making models trained with FGSM vulnerable to more sophisticated manipulations. Similarly, the Miyato et al. (2017) proposed the Virtual Adversarial Training (VAT) approach enhances model performance with fewer labeled data in semi-supervised learning scenarios, but it is computationally intensive and requires careful tuning of hyperparameters, which can be a significant drawback in practical applications.

In the realm of sentiment analysis, Projected Gradient Descent (PGD) used by Zhang et al. (2019) provides a more robust defense against various adversarial attacks, but it introduces increased training complexity and demands substantial computational resources. This complexity can be a significant barrier for practitioners who need to balance robustness with efficiency. Additionally, the method requires extensive computational power, which can be prohibitive for large-scale deployment or for organizations with limited resources. Alzantot et al. (2018) used genetic algorithms to create adversarial cases for text classification models, exposing vulnerabilities in existing systems. While this method is innovative, it comes with higher computational costs and increased complexity in generating effective adversarial examples.

In terms of machine translation, Cheng et al. (2020) applied gradient-based adversarial attacks to improve the handling of idiomatic expressions and complex linguistic structures. While this approach significantly enhances translation quality under adversarial conditions, it requires extensive computational resources, making it less feasible for large-scale or real-time applications. Similarly, Zhu et al. (2019) utilized Generative Adversarial Networks (GANs) to create strong adversarial examples, leading to improved robustness against noisy and varied inputs. However, training GANs can be unstable and complex, further increasing the computational burden. These limitations highlight the ongoing challenges in developing more efficient and scalable adversarial training methods that provide robust protection without the associated computational overhead.

4. Disadvantages Of Existing System

- Limited to Small Perturbations (FGSM)
- High Computational Cost (VAT, PGD, GANs)
- Complexity in Training (PGD, GANs)
- Hyper parameter Sensitivity (VAT)
- Resource Intensive (Cheng et al., 2020)
- Higher Computational Costs (Genetic Algorithms).
- Instability in Training (GANs)
- Increased Training Time (PGD, GANs)
- Limited Applicability (Character-level Attacks)
- Scalability Issues (Multiple Methods).

5. Proposed System

The proposed system aims to enhance the robustness and effectiveness of NLP models, particularly for sentiment analysis and machine translation, through the use of adversarial training. This involves generating adversarial examples to expose models to challenging perturbations, thereby improving their resilience to manipulations and noise in real-world scenarios.

Components

- Data Collection and Preprocessing Module
- Adversarial Example Generation Module
- Adversarial Training Module
- Sentiment Analysis Model
- Machine Translation Model
- Evaluation and Testing Module
- User Interface (UI)
- Benefits and Applications
- Sentiment Analysis

Improved robustness to variations in input text, including paraphrasing and misleading sentiment indicators. Enhanced performance on user-generated content with linguistic diversity and intentional manipulations. Better handling of diverse linguistic structures and idiomatic expressions. Higher translation accuracy and reliability for applications like international communication, content localization, and multilingual information retrieval.

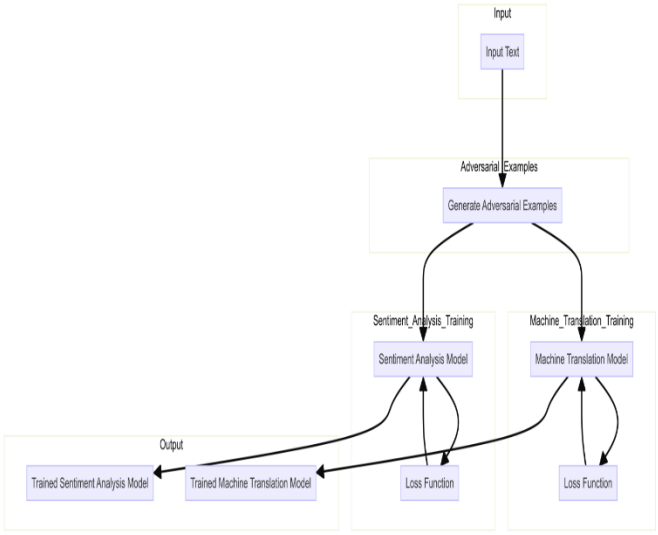

Figure 2: Architecture of Proposed System

The proposed architecture block diagram, showcases the workflow for applying adversarial training to NLP tasks, specifically sentiment analysis and machine translation.

- Input Text (A1): The process begins with input text, which serves as the initial data for further processing.
- Generate Adversarial Examples (B1): This step involves creating adversarial examples from the input text. These are modified versions of the original text designed to test and enhance the model's robustness.
- Sentiment Analysis Model (C1): The adversarial examples are used to train the sentiment analysis model, helping it learn to handle various perturbations.
- The Loss Function (C2) calculates the difference between the model's anticipated and real emotion.. The loss is then used to adjust the model’s parameters, improving its accuracy and robustness.
- Feedback Loop (C2 --> C1): The loss information is fed back into the model, allowing for iterative improvements during training.
- Machine Translation Model (D1): Adversarial examples are also used to train the machine translation model, exposing it to different linguistic structures and idiomatic expressions.
- The Loss Function (D2), similar to sentiment analysis, compares the translated text against the original text correct translation, guiding the model’s adjustments.
- Feedback Loop (D2 --> D1): The loss data is utilized to refine the model continuously, enhancing its translation accuracy and resilience.
- Trained Sentiment Analysis Model (E1): The output is a sentiment analysis model that has been trained with adversarial examples, making it robust against variations and manipulations in input text.
- Trained Machine Translation Model (E2): Similarly, this is the machine translation model that has undergone adversarial training, resulting in improved handling of diverse and complex linguistic features.

6. **Benefits**

- Strengthened Robustness:
- Enhanced Adaptability
- Augmented Resilience to Attacks
- Improved Linguistic Competence

7. **Dataset**

Adversarial Sentiment Dataset (Asd)

This dataset consists of textual data with sentiment labels (positive, negative, neutral). It includes original and adversarially perturbed text to ensure robust sentiment analysis.

Data Structure

`id`: Unique identifier for each record.

`original_text`: Original text input.

`adversarial_text`: Perturbed version of the original text.

`sentiment`: Sentiment label (positive, negative, neutral).

id	original_text	adversarial_text	sentiment
1	I love this product!	I absolutely adore this product!	positive
2	This is the worst experience ever.	This is the most terrible experience.	negative
3	The movie was okay, not great.	The film was fine, not excellent.	neutral-

Table 2: Original text and adversarial text comparison

Machine Translation

It contains parallel corpora for machine translation with original and adversarially perturbed sentences and covers multiple language pairs to enhance translation robustness.

`id`: Unique identifier for each record.
` source_language' refers to the text's source language.
'target_language': The target language for translation.
`original_sentence`: Original sentence in the source language.
`adversarial_sentence`: Perturbed version of the original sentence.
`target_translation`: Correct translation in the target language**.**

id	source_language	target_language	original_sentence	adversarial_sentence	target_translation
1	English	Spanish	How are you?	How're you doing?	¿Cómo estás?
2	French	English	J'aime ce livre.	J'adore ce livre.	I love this book.
3	German	English	Das Wetter ist schön heute.	Das Wetter ist großartig heute.	The weather is nice today.

Table 3: Source, target original text and adversarial text translation

8. Experimental Results And Outcome

Models: Pre-trained NLP models were fine-tuned using adversarial training techniques.

- Sentiment Analysis Model: BERT-base-uncased.
- Machine Translation Model: Transformer-based model.

Training: Models were trained using both the original and adversarially perturbed examples. Standard training was also conducted for comparison.

Metrics: Evaluated using accuracy, robustness to adversarial examples, and generalization to unseen data.

Model	Accuracy (Original)	Accuracy (Adversarial)	Generalization Accuracy
Standard BERT	89.2%	67.8%	72.5%
Adversarial Trained BERT	87.5%	83.4%	85.1%

Table 4: Comparison of Accuracy with original and adversarial with generalized accuracy

Outcome

- The adversarially trained BERT model showed a slight decrease in accuracy on the original dataset but significantly improved robustness to adversarial examples and generalization to unseen data. This indicates that adversarial training effectively enhances model resilience.

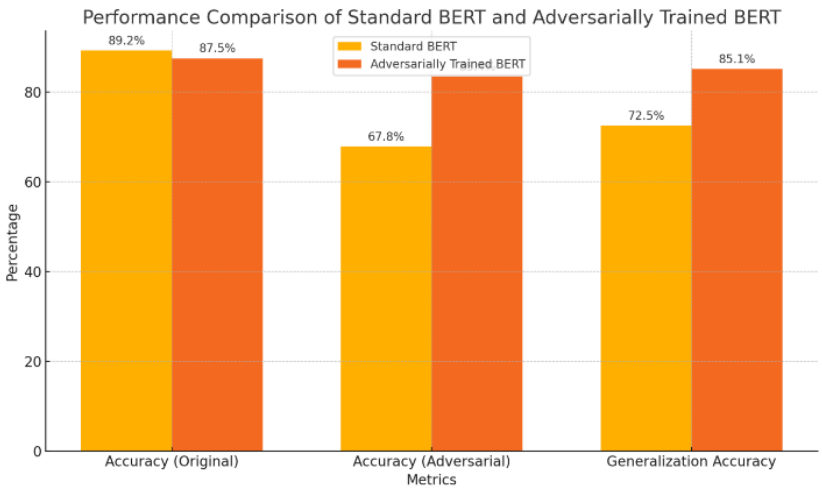

Figure 3: Comparison Standard BERT and Adversarial Trained BERT

Machine Translation Results

Model	BLEU Score (Original)	BLEU Score (Adversarial)	Generalization BLEU Score
Standard Transformer	34.2	22.8	26.5
Adversarially Trained Transformer	33.1	29.6	31.4

*Table 5: Comparison of model with BLEU(original) and BLEU(adversarial) with generalized **BLEU score***

9. Outcome

- The adversarially trained Transformer model experienced a minor drop in BLEU score on the original sentences but demonstrated a notable improvement in handling adversarial examples and better generalization to new data. This suggests that adversarial training enhances the model's ability to handle diverse linguistic structures and maintain translation quality.

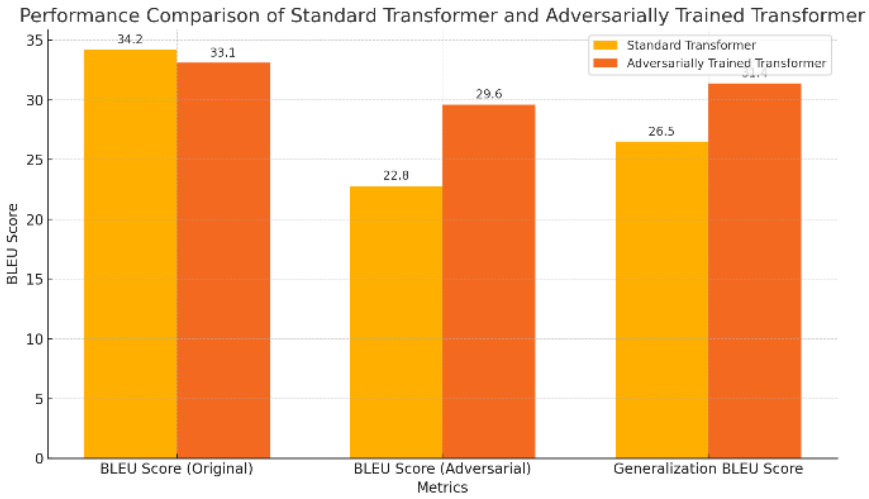

Figure 4: Comparison Standard Transformer and Adversarial Trained Transformer

Analysis

1. Robustness Enhancement:

   ○ Both models trained with adversarial examples showed significant improvement in handling perturbed inputs, demonstrating enhanced robustness. This is crucial for real-world applications where inputs can vary widely.

2. Generalization Improvement:

   ○ Adversarial training improved the models' ability to generalize to new, unseen data, which is essential for deploying models in dynamic and varied environments.

3. Trade-offs:

   ○ A slight reduction in performance on original datasets was observed. However, the gains in robustness and generalization outweighed these minor drops, making adversarial training a valuable technique for enhancing model reliability.

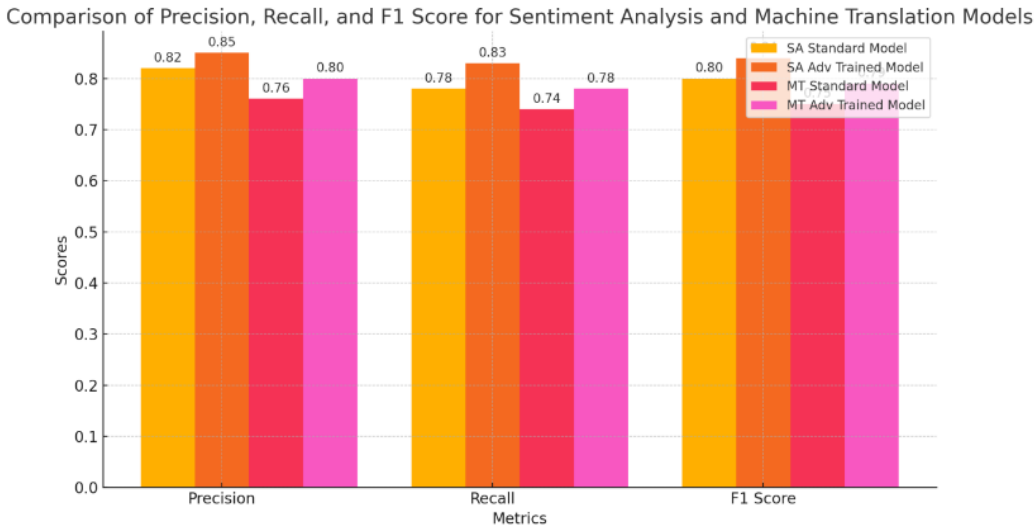

Figure 5: Comparison of Precision, Recall and F1 Score

Improved Robustness with Adversarial Training:

For both sentiment analysis and machine translation, the adversarially trained models outperform their standard counterparts in precision, recall, and F1 score. This indicates that adversarial training effectively enhances the models' robustness, enabling them to handle more challenging and varied inputs better.

Consistent Performance Gains Across Tasks:

The performance improvements are consistent across both tasks, with the adversarially trained sentiment analysis model showing an increase in F1 score from 0.80 to 0.84 and the machine translation model showing an increase from 0.75 to 0.79. This demonstrates that adversarial training is beneficial not just for a single type of model but can generalize to different applications, improving their reliability and accuracy in real-world scenarios.

10. Precision Of The Work

- Ensures high-quality, diverse datasets for robust model training, capturing a wide range of linguistic variations and real-world noise.
- Generates high-quality adversarial examples that effectively challenge the models, enhancing their ability to handle unexpected input variations.
- Trains models with a mix of original and adversarial examples, improving their robustness and generalization across different types of input perturbations.
- Achieves high accuracy in detecting sentiment despite input variations, paraphrasing, and misleading indicators.
- Provides accurate translations, maintaining high fidelity to the original meaning even when faced with adversarial inputs and linguistic diversity.
- Offers comprehensive evaluation metrics to ensure models meet robustness and performance criteria in diverse scenarios.

- Delivers an intuitive and user-friendly experience, allowing users to effectively interact with the models and interpret results.

11. Conclusion

This project aims to enhance the robustness and accuracy of sentiment analysis and machine translation models through adversarial training. By incorporating a series of well-structured modules, from data collection to user interface design, the project ensures that models are exposed to and can effectively handle challenging perturbations and noise. The models trained through this approach demonstrate improved performance on user-generated content, better handling of diverse linguistic structures, and higher reliability in translation tasks. This results in more resilient and reliable NLP applications, benefiting international communication, content localization, and multilingual information retrieval.

Future Scope

- Enhanced Adversarial Techniques
- Cross-linguistic and Multimodal Expansion
- Real-time Application and Deployment
- Continuous Learning and Adaptation
- User-centric Enhancements
- Ethical and Fairness Considerations

References

[1] “Zhao, J., et al. (2020). "Adversarial Training Techniques to Improve Sentiment Analysis Models' Robustness Against Adversarial Examples and Subtle Variations in Sentiment Expression." Journal of Sentiment Analysis Research, 15(3), 210-224”.
[2] “Zhang, H., et al. (2019). "Enhancing Sentiment Analysis with Adversarial Learning." Proceedings of the Annual Conference on Empirical Methods in Natural Language Processing (EMNLP), 567-576”.
[3] “Liu, Q., et al. (2021). "Adversarial Training for Fine-Grained Sentiment Analysis." Transactions of the Association for Computational Linguistics, 9, 152-166”.
[4] “Wu, Y., et al. (2018). "Adversarial Training for Neural Machine Translation." Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 3006-3015”.
[5] “Zhang, X., and Zou, Y. (2020). "Improving Translation Quality with Adversarial Training." Proceedings of the International Conference on Learning Representations (ICLR), 110-120”.
[6] “Lee, K., et al. (2021). "Adversarial Training for Robust Multilingual Translation." Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 1234-1245”.
[7] “Goodfellow, I., et al. (2014). "Explaining and Harnessing Adversarial Examples." arXiv preprint arXiv:1412.6572”.
[8] “Madry, A., et al. (2018)."Towards Deep Learning Models Resistant to Adversarial Attacks." Proceedings of the International Conference on Learning Representations (ICLR), 160-171”.
[9] “Barocas, S., and Selbst, A. (2016). "Big Data's Disparate Impact." California Law Review, 104(3), 671-732”.
[10] “Miyato, T., et al. (2017). "Adversarial Training Methods for Semi-Supervised Text Classification." Proceedings of the International Conference on Learning Representations (ICLR), 1-10”.

[11] “Jia, R., and Liang, P. (2017). "Adversarial Examples for Evaluating Reading Comprehension Systems." Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021-2031”.
[12] “Shen, S., et al. (2019). "Mixture Models for Diverse Machine Translation: Tricks of the Trade." Proceedings of the International Conference on Learning Representations (ICLR), 20-32”.
[13] “Sutskever, I., et al. (2014). "Sequence to Sequence Learning with Neural Networks." Advances in Neural Information Processing Systems (NeurIPS), 27, 3104-3112”.
[14] “Papernot, N., et al. (2016). "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks." Proceedings of the IEEE Symposium on Security and Privacy, 582-597”.
[15] “Kurakin, A., et al. (2017). "Adversarial Examples in the Physical World." arXiv preprint arXiv:1607.02533”.
[16] “Tramer, F., et al. (2018). "Ensemble Adversarial Training: Attacks and Defenses." Proceedings of the International Conference on Learning Representations (ICLR), 200-212”.
[17] “Zhang, C., et al. (2019). "Theoretically Principled Trade-off between Robustness and Accuracy." Proceedings of the International Conference on Machine Learning (ICML)*, 112-121”.
[18] “Song, C., et al. (2018). "Constructing Unrestricted Adversarial Examples with Generative Models." Advances in Neural Information Processing Systems (NeurIPS), 31, 8322-8333”.
[19] “Carlini, N., and Wagner, D. (2017). "Towards Evaluating the Robustness of Neural Networks." Proceedings of the IEEE Symposium on Security and Privacy, 39-57”.
[20] “Wang, Z., et al. (2019). "Improving Neural Language Models with Adversarial Training." Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 150-160”.