# Attack methods on LSTM-Based Recurrent Neural Networks for Sentiment Analysis

Rafał Zan

Faculty of Electrical Engineering
Warsaw University of Technology

## Introduction

Recurrent Neural Networks (RNN), especially their variants like LSTM (Long Short-Term Memory), have become a key tool in natural language processing (NLP). Their ability to effectively capture sequential dependencies in text data makes them ideal for sentiment analysis in user-generated content (e.g., posts, reviews). Despite their effectiveness, these models are susceptible to deliberate disturbances in input data (adversarial attacks), which can disrupt their classification capabilities.

## Research Objective

This work aims to analyze the impact of input manipulations on LSTM networks and compare their resilience with classical RNN networks in the sentiment analysis task. The research was conducted on specially prepared RNN and LSTM models, trained on a dataset containing game reviews from Twitter. The impact of two types of attacks on classification results was evaluated.

## Models and Data

- **Models:** Standard Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks were compared.
- **Task:** Sentiment Analysis (text sentiment classification).
- **Data:** Custom dataset of game reviews collected from Twitter. The models were trained to recognize positive or negative sentiment in these reviews.

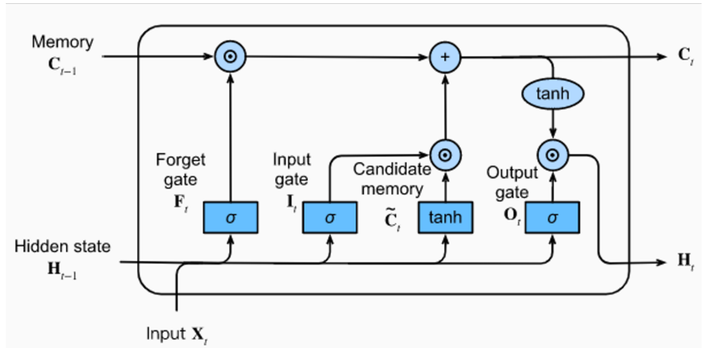


Figure 1. LSTM cell architecture.

## Adversarial Attack Methods

Two techniques for modifying input data aimed at reducing model effectiveness were analyzed:

- **Synonymization:** Replacing words in the text with their synonyms.
- **Token Replacement with Similar Vectors:** Replacing words (tokens) with other words whose vector representations (embeddings) are most similar.

The goal of both attacks is to change the model's prediction with minimal modification to the original text.

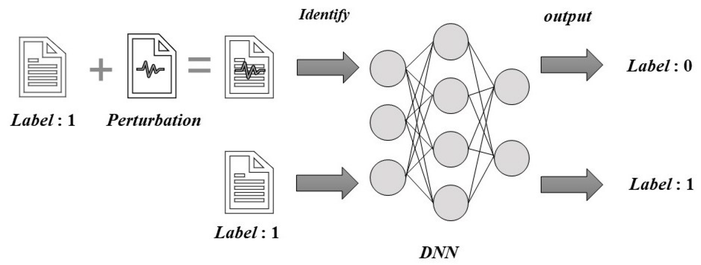


Figure 2. Attack process schema.

## Experimental Procedure

1. RNN and LSTM models were trained on the prepared review dataset.
2. Test samples were selected and both attack methods were applied to them.
3. The effectiveness of models on original and modified samples was evaluated.
4. The resilience of the LSTM model and standard RNN to both types of attacks was compared.

Example of text that was disturbed using the PGD method:

| | |
|---|---|
| Original Text | mentioned facebook struggling motivation go run day translated toms great auntie hayley cant get bed told grandma thinks im lazy terrible person |
| Perturbed Text | relax tweets relax relax visit complaining ears release horse horse rockstar callofduty callofduty sum sum sum lowe annoying passed passed |

Table 1. Original and Perturbed text comparison

## Results

The conducted experiments showed that:

- **LSTM models demonstrate greater resilience** to synonymization compared to standard RNN networks.
- **Both types of networks, including LSTM, remain susceptible** to attacks based on replacing tokens with semantically similar vectors.
- Creating effective adversarial examples for text is **feasible**, although more complex than for images.

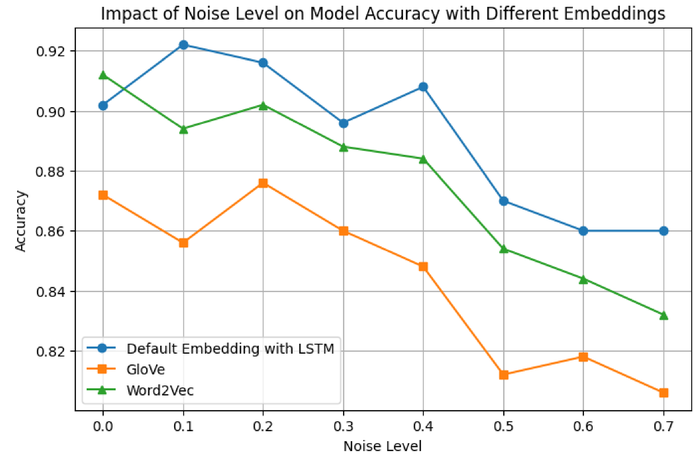


Figure 3. Comparison of model accuracy before and after attack.

## Conclusions and Future Directions

The research confirmed the vulnerability of NLP models to adversarial attacks. Further work on detection and defense methods, including adversarial training [2], is necessary. AI model security is crucial in the context of their growing popularity.

## References

[1] Mark Anderson, Andrew Bartolo, and Pulkit Tandon. Crafting adversarial attacks on recurrent neural networks. n.d., n.d. Accessed [Your Access Date].

[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, April 2021. doi: 10.48550/arXiv.2102.01356. URL `https://doi.org/10.48550/arXiv.2102.01356`.

[3] Simon Geisler, Tom Wollschläger, M. H. I. Abdalla, Johannes Gasteiger, and Stephan Günnemann. Attacking large language models with projected gradient descent. *arXiv preprint arXiv:2402.09154*, February 2024. doi: 10.48550/arXiv.2402.09154. URL `https://doi.org/10.48550/arXiv.2402.09154`.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, March 2015. doi: 10.48550/arXiv.1412.6572. URL `https://doi.org/10.48550/arXiv.1412.6572`.

ORCID: 0009-0001-2574-0887



rafal.zan.stud@pw.edu.pl