Growth Bound Matrix and Adversarial Training for Robustness in S4 Text Classification

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the Growth Bound Matrix approach compare to adversarial training in improving the robustness of S4 models against synonym substitution attacks on text classification benchmarks. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Bridging Robustness and Generalization Against Word Substitution Attacks in NLP via the Growth Bound Matrix Approach. Research question: How does the Growth Bound Matrix approach compare to adversarial training in improving the robustness of S4 models against synonym substitution attacks on text classification benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

14 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The IMDB dataset contains 25,000 movie reviews for training and 25,000 for testing.	×	0.01
The Yahoo! Answers dataset contains 1,400,000 training samples and 50,000 testing samples across 10 classes.	×	0.05
Adversarial examples were generated using 1000 randomly sampled test instances from each dataset.	×	0.06
For comparison with Text-CRS, 500 instances were sampled from the IMDB test set to assess robustness against the TextFoo	×	0.04
The perturbation setting uses k=8 nearest neighbors within a Euclidean distance of de=0.5 in the GloVe embedding space.	×	0.03
On the IMDB dataset with a CNN model, the Standard defense achieves 89.7% clean accuracy and 0.6% accuracy under PWWS at	×	0.07
On the IMDB dataset with a CNN model, the GBM defense achieves 90.2% clean accuracy and 76.3% accuracy under PWWS attack	×	0.04
On the IMDB dataset with a BiLSTM model, the GBM defense achieves 84.3% accuracy under PSO attack, compared to 75.5% for	×	0.05
On the Yahoo! Answers dataset with a CNN model, the Standard defense achieves 72.6% clean accuracy and 6.8% accuracy und	×	0.07
On the Yahoo! Answers dataset with a BiLSTM model, the GBM defense achieves 68.6% accuracy under PSO attack.	×	0.03
Training the GBM approach on a BiLSTM model for the IMDB dataset takes 25 seconds per epoch.	×	0.03
Training the IBP approach on a BiLSTM model for the IMDB dataset takes 4 minutes and 53 seconds per epoch.	×	0.02
The GBM approach reduces training time by more than 11 times compared to IBP.	×	0.03
The GBM training method for the BiLSTM model uses an Adam optimizer with a batch size of 64 and a learning rate of 10^-3	×	0.03
The maximum sequence length used for the IMDB dataset is 512, and for Yahoo! Answers is 256.	×	0.02

References

- http://arxiv.org/abs/2507.10330v1
- http://arxiv.org/abs/2008.03709v4
- http://arxiv.org/abs/2106.01065v2