SOVEREIGN: How does fine-tuning Llama3, Codestral, and Deepseek R1 on the full Big-Vul dataset impact their vulnerability

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

With the advent of the modern pre-trained Transformers, the text preprocessing has started to be neglected and not specificly addressed in recent NLP literature. However, both from a linguistic and from a computer science point of view, we believe that even when using modern Transformers, text preprocessing can significantly impact on the performance of a classification model. We want to investigate and compare, through this study, how preprocessing impacts on the Text Classification (TC) performance of modern and traditional classification models. We report and discuss the preprocessing techn

1 Introduction

Analysis of: Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Research goal: How does fine-tuning Llama3, Codestral, and Deepseek R1 on the full Big-Vul dataset impact their vulnerability classification accuracy and consistency across different programming languages?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

4 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 8.3/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
The advent of modern pre-trained Transformers has led to text preprocessing being neglected and not specifically address	✓	0.26
Text preprocessing can significantly impact the performance of a classification model, even when using modern Transforme	✓	0.33
The study investigates and compares how preprocessing impacts the Text Classification (TC) performance of modern and tra	✓	0.27
The study reports and discusses preprocessing techniques found in the literature and their most recent variants or appli	✓	0.28
The study applies the three top referenced preprocessing techniques (alone or in combination) to four publicly available	✓	0.26
Nine machine learning models, including modern Transformers, are used with preprocessed text as input.	✓	0.24
Choosing the best preprocessing technique over the worst can significantly improve accuracy on the classification, up to	✓	0.31
In some cases, a suitable preprocessing strategy can make a significant difference in classification performance.	✓	0.15

References

- https://doi.org/10.1016/j.is.2023.102342

- https://doi.org/10.1038/s41746-022-00589-7
- https://doi.org/10.1186/s42400-025-00361-w