Necessity and Sufficiency for Explaining Text Classifiers
SNLP 2025 Project Plan

Valentin Dorseuil, Thomas Robert, Imen Waffra, Mouad Id Sougou

February 14, 2025

1 Introduction

As machine learning models become widely used and adopted, it is crucial to understand how they make decisions. However, many models act as black boxes, making it difficult to explain their prediction.

While classical Explainable AI (XAI) techniques, such as feature importance scores or attention heatmaps, can offer some insights, they often fall short in NLP tasks, where context, word relationships, and language nuances play a crucial role.

In this context, the paper [1] introduces an explainability framework based on **necessity and sufficiency** scores. **Necessity** reflects how essential a word is for the model's prediction, while **sufficiency** indicates whether the word alone can produce the same output. Applied to hate speech classifiers, this approach highlights biased patterns, such as false positives caused by identity-related words (e.g., "Muslim," "women"), exposing potential sources of biases and unfair outcomes.

2 Project ambitions

After replicating the main result of the paper, we will be exploring the following three main ideas :

Exploration, Descriptive statistics We will first conduct an exploratory study and compute descriptive statistics. The paper introduces a novel feature attribution method for text classifiers, distinguishing between necessity and sufficiency to offer more informative explanations. We will explore some datasets, understand the distribution of hate speech examples, and evaluate the classifier's reliance on identity terms. We aim to identify potential biases in the classifier with the necessity and sufficiency metrics.

Trade-off between complexity and social bias By definition, hate speech is directed either at an individual or at a community. For example, a hate speech classifier might rely on social biases toward a particular community to identify hate speech, focusing on the presence of trigger tokens rather than achieving a semantic understanding of the text. The introduction of necessity and sufficiency metrics helps to identify such biases. Using these methods, we aim to compare classifier training frameworks to identify those that tend to over-rely on specific token utterances for classification. Subject to computational complexity, we aim to: (1) compare models of different sizes to answer the question: do larger models better understand sentence semantics for hate speech detection?; and (2) compare fine-tuning methods to answer the question: do parameter-efficient fine-tuning (PEFT) methods, such as low-rank adaptation (LoRA), lead to poorer semantic classification?

Debiase the classifier Our final approach for this project involves using the necessity and sufficiency values generated by the classifier to fine-tune it towards an unbiased state. For instance, one could argue that words like *muslims* and *christians* should have identical necessity and sufficiency scores if the classifier is unbiased. By doing so, we aim to enhance the fairness of the classifier, which can sometimes be overly sensitive to identity terms.

3 Collaboration

We will be organized in the following: Imen will focus on computing statistics for various models and datasets to identify biased or unbiased datasets. Mouad and Thomas will be responsible for fine-tuning the generative model. Meanwhile, Valentin will explore using these results to fine-tune an unbiased classifier. We plan to collaborate frequently on Thursday following the lecture and share code via GitHub.

References

[1] Esma Balkir et al. *Necessity and Sufficiency for Explaining Text Classifiers: A Case Study in Hate Speech Detection*. 2022. arXiv: 2205.03302 [cs.CL]. URL: https://arxiv.org/abs/2205.03302.