Hate Speech Detection - Project Proposal - NLP

Căbulea Victor-Andrei, AI 1A

Flămînzanu-Mateiuc Marian, AI 1A

Project Idea

This project focuses on developing a system capable of detecting hate speech in textual content using transfer learning with large pre-trained language models. The primary objective is to build and fine-tune a binary text classifier that can distinguish between hate speech and non-hateful content.

As online platforms continue to grow, so does the usage of toxic and harmful language. Manual moderation is neither scalable nor consistent, and traditional machine learning techniques often fall short in capturing the nuanced and context-dependent nature of hate speech. Thus, there is a necessity to have intelligent solutions to content moderation challenges. By using models like BERT and its domain-specific variant HateBERT, we aim to build a classifier that understands subtle semantic cues and has good generalization results.

Technical Approach

The approach involves fine-tuning a pre-trained transformer-based language model on a dataset combining multiple hate speech corpora. Currently, HateBERT—a version of BERT pre-trained on Reddit hate speech data—is the model of choice due to its domain-specific strengths. To train and evaluate the classifier, we will use labeled data from sources such as HateXplain and the Davidson dataset.

The evaluation strategy will focus on key metrics like the macro-averaged F1 score, precision, recall, and confusion matrix, which will help us assess the model's ability to correctly identify both hate and non-hate content. Class imbalance will be handled through techniques such as weighted sampling to ensure fairness across categories.

Expectations

We expect the fine-tuned model to achieve competitive performance on validation data, with a macro F1 score exceeding 0.80. Additionally, the model should demonstrate strong generalization across multiple hate speech sources. Beyond accuracy, we aim for the system to be explainable, reproducible, and practically usable, paving the way for future extensions in multilingual or multi-label hate speech detection.