

Python Project: Anti-Hater Filter for Social Networks

Project Description:

I worked on a Deep Learning based “Anti-Hate Filter” system to automatically moderate user comments on a tech forum. The goal was to detect multiple types of toxic content (threats, insults, obscene language, identity-hate etc.) and classify each comment into one or more toxicity labels, improving safety and quality of online discussions.

Key Responsibilities and Tasks

I prepared and cleaned the dataset of ~160k comments, tokenized the text and balanced the labels. I designed and trained a recurrent neural network architecture (LSTM/GRU) for multi-label classification, implemented model optimization strategies, and evaluated performance through accuracy, F1-score and per-category metrics. The system outputs a vector of 6 binary values, one for each toxicity category, to be used for real-time moderation.

Outcome

This solution significantly reduced the moderation workload, enabling automatic screening of toxic comments without manual review. Thanks to recurrent layers capturing context, the model increased detection accuracy compared to rule-based or traditional approaches, while being scalable to growing traffic.

Technologies and Tools Used

TensorFlow / Keras (RNNs), Pandas, NLTK / tokenizers, deep learning for NLP, multi-label classification.