Hate Speech Detection Using Natural Language Processing and Deep Learning

Soluyanova Zlata, Semenova Diana

Proposal Idea

The proposal idea: a structured project to identify hate speech using BiLSTM

Identifying hate speech online remains challenging due to contextual nuances such as sarcasm, coded language, and hidden toxicity. Traditional models (for example, logistic regression) often fail to capture sequential dependencies, which leads to poor generalization. This project addresses these gaps by focusing on contextual identification to improve accuracy and interpretability in real-world moderation systems.

We will use Kaggle's hate speech dataset (https://www.kaggle.com/datasets/waalbannyantudre/hate-speech-detection-curated-dataset?resource=download), a carefully selected collection of texts labeled offensive or obscene. The dataset includes diverse examples of hate speech, slang, and hidden toxicity. Preprocessing steps include tokenization, lemmatization, and handling class imbalance (e.g., via focal loss), ensuring robust training for minority classes.

To solve the problem, we plan to use bidirectional LSTM (BiLSTM) with an attention layer. BiLSTM processes text bidirectionally, capturing context from past and future words. GloVe embeddings (vector representations of words) encode semantic connections, while the attention-grabbing mechanism highlights discriminatory phrases, improving interpretability.

Model performance will be rigorously assessed using precision, recall, and F1-score to account for class imbalance in hate speech detection. We will compare our BiLSTM model against baseline – logistic regression (for its simplicity and interpretability).