Link to Presentation:
https://docs.google.com/presentation/d/1sjVvxzqwskSXf9s6js6hAK8bihosDCRRn19t7aAfY-o/edit?slide=id.p10#slide=id.p10

Writeup

For the strong baseline in Milestone 2, we fine-tuned a BERT model on the MetaHate dataset to detect hate speech. BERT provides powerful contextual embeddings and effectively captures relationships between words across a sentence, but it has certain limitations. While BERT incorporates positional encodings to represent the sequential positions of tokens, its self-attention mechanism is inherently permutation-invariant, meaning that, irrespective of token order, it considers all token relationships equivalently, potentially yielding similar outputs. Consequently, changes in word order, such as "They are allies" versus "Allies they are," may not be fully distinguished, even though the first is neutral or positive, while the second could be interpreted sarcastically or as hostile depending on context. To address these limitations, we proposed two model extensions: BERT + CNN and BERT + LSTM.

The first extension, BERT + CNN, aims to capture local n-gram patterns from BERT's embeddings, such as short phrases like "Go back to ___", "I can't stand those ___", or "Ban all ___" that are strong indicators of hateful content. By applying convolutional layers on top of BERT embeddings, the model can detect these local patterns while still leveraging BERT's contextual understanding. This combination allows the model to be sensitive to both global context and meaningful local word combinations, leading to improved detection of hate speech signals. The BERT + CNN model achieved an F1 score of 0.706.

The second extension, BERT + LSTM, explicitly models the sequential structure of words, enhancing sensitivity to word order and helping BERT compensate for self-attention's permutation-invariance. By using an LSTM on top of BERT embeddings, the model can better capture sequential dependencies, including long-range interactions between words and phrases where rearranging words changes meaning. This is especially useful for handling negations, sarcasm, or multi-clause statements often found in hate speech. The BERT + LSTM model achieved an F1 score of 0.696.