

Milestone 1

Dataset Description: For our project on hate speech detection, we use the MetaHate dataset, which is a meta-collection of 36 hate speech datasets compiled from social media comments. The dataset contains 1,226,202 posts, of which 1,101,165 instances are publicly available, and each post is labeled for hate speech. The labels are binary: 0 indicates non-hate speech, while 1 indicates hate speech. The dataset is provided in TSV (tab-separated values) format, making it easy to load and process for NLP tasks. The MetaHate dataset is publicly available on [Hugging Face](#), and the accompanying [paper](#) provides further details about its compilation and usage. We chose this dataset because it is highly relevant to our hate speech classification task, as it contains social media posts labeled as either hate speech or non-hate speech.

BERT-based Ensemble Approaches for Hate Speech Detection: This paper investigates how combining transformer-based models and neural networks can improve the automated detection of hate speech on social media platforms such as Twitter. The authors extend the capabilities of BERT (Bidirectional Encoder Representations from Transformers) by integrating it with three different neural architectures, a Multi-Layer Perceptron (MLP), a Convolutional Neural Network (CNN), and a Long Short-Term Memory (LSTM) network, to capture complementary linguistic features. Using three publicly available Twitter datasets (Davidson, HatEval2019, and OLID), they created a unified, more balanced dataset called DHO, designed for multi-label classification of hateful, offensive, and neutral content. To enhance model performance and robustness, the study employed four ensemble learning methods: (1) Soft voting, where each model outputs class probabilities and the final prediction is the class with the highest *average* probability across models. (2) Max voting, which selects the class predicted with the highest individual probability among all models. (3) Hard voting, which uses the *majority vote* among the models' discrete predictions (e.g., if two out of three predict "offensive," that becomes the final label). (4) Stacking (stacked generalization), a more advanced technique where predictions from base models (e.g., BERT+MLP, BERT+CNN, BERT+LSTM) are fed into a meta-classifier, in this case, a linear regression model, that learns to optimally combine them for the final output. The experiments showed that, for the most part, BERT+NN hybrid models outperformed the plain BERT baseline, highlighting the benefit of combining contextual embeddings with neural feature extractors. Ensemble techniques further improved performance: on the Davidson dataset, the stacking ensemble achieved a remarkable 97% F1-score, surpassing other ensemble methods. On the larger DHO dataset, ensembles involving BERT+MLP and BERT+LSTM achieved the best macro-F1 and precision scores, though at a higher computational cost. Overall, the paper demonstrates that combining transfer learning (BERT) with ensemble learning yields superior hate speech detection performance, capturing subtle and context-dependent language. The authors note that future work should address class imbalance, computational efficiency, and interpretability, potentially through architectures like K-BERT or enhanced ensemble framework.

HateBERT: Retraining BERT for Abusive Language Detection in English: This paper from Caselli et al., 2021 introduces HateBERT, a version of BERT specially retrained to improve the

detection of abusive and hateful language on social media. The authors aim to make BERT more sensitive to toxic and aggressive forms of online communication by adapting it to the domain of abusive language. To do this, they used domain-specific fine-tuning, where the standard English BERT model (base uncased) is further trained on language drawn from online communities known for toxic behavior. The new training data, called RAL-E, contains over one million Reddit comments collected from communities that were banned for promoting offensive, abusive, or hateful content. By fine-tuning the model to this kind of text, the authors tried to shift BERT’s understanding of language toward the patterns and tone found in abusive speech. The team compared the retrained HateBERT model with the original BERT across three benchmark datasets that focus on different aspects of harmful online language: OffensEval 2019 for offensive language, AbusEval for abusive language, and HatEval for hate speech. In all cases, HateBERT achieved higher macro-averaged F1 scores, showing that it was better at identifying both harmful and non-harmful messages. The biggest improvement appeared on the AbusEval dataset, where HateBERT even surpassed previous best results. The authors also tested how well the models transferred between datasets. These portability experiments showed that HateBERT performed more reliably across related tasks, especially when moving from broader categories like offensive language to more specific ones like hate speech. However, they also found that the ability to generalize depends on how similar the labeled categories are between datasets. Overall, the paper shows that further fine-tuning of LLMs on domain-specific data is an effective and low-cost way to adapt them to specialized tasks such as hate speech detection. The authors released both the HateBERT model and the RAL-E dataset to support future research.

Advancing Hate Speech Detection with Transformers: Insights from the MetaHate: The paper presents a large-scale study on transformer-based models for hate speech detection using the previously introduced MetaHate dataset. The authors begin by outlining the limitations of traditional machine learning and deep learning approaches—such as SVMs, CNNs, and LSTMs—in capturing the complex contextual nuances of hate speech. They then systematically evaluate several transformer architectures, including BERT, RoBERTa, GPT2, BART, DeBERTa, Longformer, XLNet, T5, and ELECTRA, to identify which model most effectively detects hate speech across diverse and noisy social media data. The dataset was tokenized, balanced using class weights, and all models were fine-tuned under consistent hyperparameter settings to ensure fair comparison. Among all evaluated models, **ELECTRA** achieved the best performance with an F1 score of **0.8980** and accuracy of **0.8946**, outperforming all baselines as well as other transformer architectures. ELECTRA’s generator–discriminator framework—where the generator replaces masked tokens and the discriminator identifies those replacements—enables it to capture subtle contextual cues, making it especially effective at detecting implicit or coded hate speech. Error analysis revealed ongoing challenges with sarcasm, figurative language, and label noise. Overall, the study demonstrates that transformer-based architectures, particularly ELECTRA, substantially advance hate speech detection, while future work should focus on explainable and multimodal extensions to enhance robustness and fairness.