Hate Speech Detection with LLM

Students

Lukas Derungs

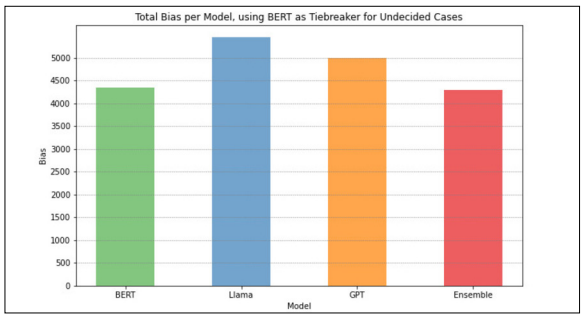
Kailing Peng

Introduction: The rise of online hate speech, exacerbated by social media, impacts psychological health and can incite violence. Effective detection methods are crucial to prevent the spread of such content. Traditional detection methods using lexicons struggle to keep up with the evolving, context-dependent nature of hate speech, while fine-tuning (further training a model to excel in a specific field) is commonly done with datasets targeting specific groups, e.g., women, LGBTQ. In recent years, Large Language Models (LLMs), which demonstrate human-like thinking and analysis, have been employed for hate speech detection to address these challenges.

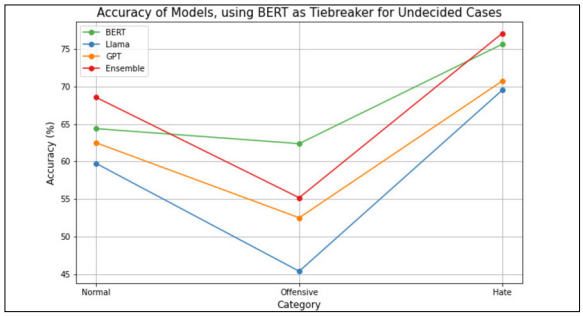
Definition of Task: Bias is a prevalent issue in pre-trained LLMs, requiring task-specific bias identification and mitigation. Our study examines bias in LLMs by experimenting with text-based hate speech detection using In-Context Learning (ICL, providing a few correct examples in the prompt for the model to learn from) for GPT-4.0 mini and Llama 3.1 8B, alongside a fine-tuned BERT model. We formed a voting ensemble to assess if majority voting can balance individual model biases and improve categorization accuracy for hate speech, offensive language, and normal language.

Conclusion: Our results demonstrate that while all the LLMs we used exhibit different biases towards various labels in the dataset (e.g., normal language being classified as offensive), majority voting can effectively reduces bias and improves accuracy when classifying hate speech and normal language compared to each individual model. However, in scenarios where the performance of the participating models vary drastically, particularly in the category offensive, the ensemble approach does not outperform the best single model. Furthermore, the voting ensemble encountered cases where a draw occurred, comprising about 5.4% of the total 13,229 data entries. Separate considerations were made for these ambiguous cases—either removing them from the evaluation or replacing them with the best model's decision. Here, we present the version using the best model to represent the majority when there is a tie.

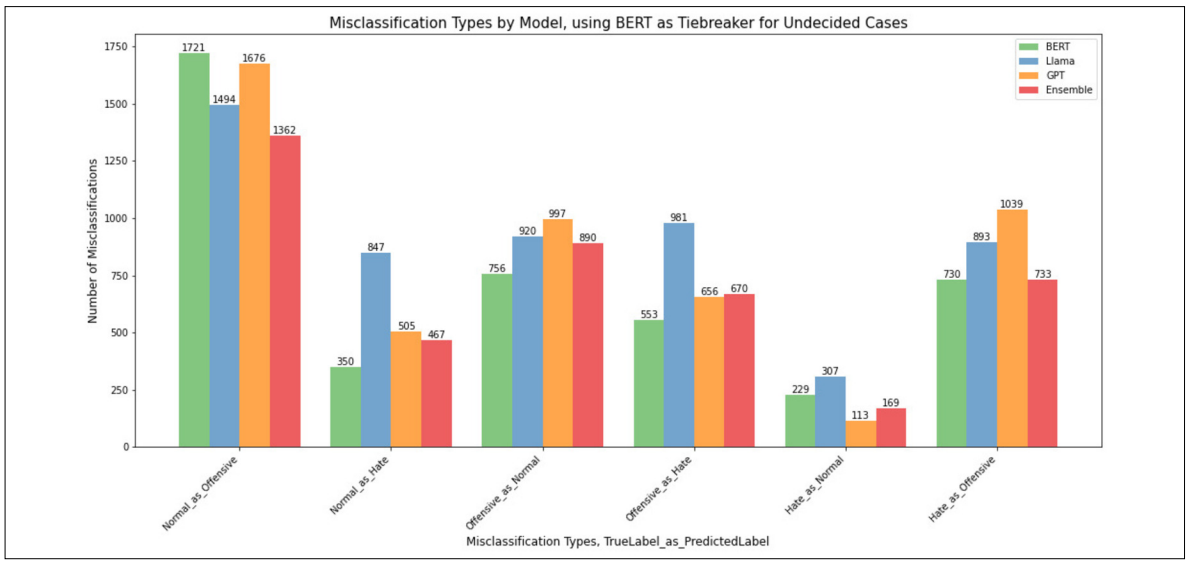
Overview of the Total Bias in a Given Model
Own presentment



Model Accuracy per Category
Own presentment



Misclassification Types Compared to True Labels
Own presentment



Advisor
Prof. Dr. Daniel Patrick Politze

Subject Area
Artificial Intelligence, Miscellaneous