**Description of Evaluation Measure**

This project focuses on leveraging state-of-the-art (SOTA) NLP techniques for hate speech detection using the MetaHate dataset, a unified benchmark for this task. Specifically, given a social media post, the goal is to classify it as hateful or non-hateful. Since this is a binary classification problem, the F1 score will be used as the primary evaluation metric, as it balances Precision and Recall—two critical aspects in hate speech detection. Recall captures the proportion of actual hate speech posts correctly identified by the model. Optimizing for recall ensures harmful content is not overlooked. Precision measures the proportion of posts labeled as hate speech that are truly hateful. Optimizing for precision reduces the risk of mislabeling benign content and unnecessarily restricting users' expression. For social media platforms represented in the MetaHate dataset, both metrics are essential: missing hate speech allows harmful content to spread, while false positives can unfairly penalize users. Striking the right balance ensures the platform remains safe and inclusive without over-censoring legitimate expression, maintaining trust and engagement among users. In this work, the F1 score is computed specifically for the hate speech class (labeled as 1 in our dataset), as correctly identifying hateful content is our primary concern.

Mathematically, Precision, Recall, and F1 score are defined as follows: Precision = $\frac{TP}{TP + FP}$, Recall = $\frac{TP}{TP + FN}$, and F1 score = $\frac{2 * Precision * Recall}{Precision + Recall}$, where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. The F1 score provides a single metric that balances Precision and Recall, making it particularly suitable for tasks like hate speech detection where both false negatives and false positives carry significant consequences.

We want to note that accuracy alone is insufficient in this scenario because the dataset is imbalanced, with only about 20% of posts labeled as hate speech. A naive classifier that always predicts "non-hateful" could achieve 80% accuracy without detecting any actual hate speech. By using the F1 score as the primary metric, we ensure that the model's performance reflects its ability to correctly identify hateful content while minimizing false positives, providing a more meaningful evaluation for this imbalanced dataset.

To conclude, we note that our strong baseline draws direct inspiration from the paper MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection, which fine-tunes a BERT model on the MetaHate dataset. In their experiments, they also use F1 score as an evaluation metric, emphasizing its importance for balancing Precision and Recall in hate speech detection tasks.

**Simple Baseline Description and Performance:**

This baseline model is a simple majority-class classifier that ignores all text features and relies solely on the label distribution in the training set. During training, this simple baseline identifies the single most frequent class in the training set between 0 (non-hate) and 1 (hate). This majority class becomes the model's single prediction rule. On new unseen data, the model consistently

predicts this class for every example. Because it does not analyze text or learn any linguistic patterns, this baseline sets the lowest meaningful performance threshold that more advanced models, such as BERT, are expected to exceed. It provides a naive benchmark for assessing the impact of class imbalance and for determining whether stronger models truly learn features associated with hateful speech, rather than simply leveraging label priors from the training data. When evaluating this simple baseline on our test dataset, we observed the following results:

- **F1 Score: 0.0**

The F1 score of 0.0 for this baseline arises because it always predicts the majority class, which in our dataset is non-hate (0). Since only 20% of the posts are actually hate speech (1), the model never predicts any true positives. In practice, this is horrible because the model completely fails to identify any hate speech. While it may achieve 80% accuracy by predicting only non-hate (matching the 80% majority class), it is utterly useless for detecting harmful content, which is the main goal. This goes to further illustrate why accuracy is misleading on imbalanced datasets and why more sophisticated models that actually learn textual patterns are essential.

**Strong Baseline Description and Performance**

This strong baseline leverages a fine-tuned BERT model (bert-base-uncased) for hate speech detection. Unlike a simple majority-class classifier, this model benefits from the rich contextual language representations learned during BERT's pre-training and adapts them to the MetaHate dataset via supervised fine-tuning. All social media posts in the MetaHate dataset are tokenized using BERT's tokenizer with truncation and padding to a maximum sequence length of 512 tokens, enabling batch processing.

To address the class imbalance in the dataset—where hateful posts make up only about 20% of the training data while non-hateful posts constitute the remaining 80%—we employ a class-weighted cross-entropy loss. The weights are determined based on the proportion of each class in the training data, so that the model gives more emphasis to the minority class (hateful posts) during training. This custom loss is implemented by subclassing Hugging Face's Trainer class, which allows the model to learn effectively from both classes while avoiding being biased toward the majority class.

We fine-tuned the BERT model for three epochs using a batch size of 32 and a learning rate of 5e-5, with the AdamW optimizer for parameter updates. Evaluation focuses on the F1 score for the hate speech class (label = 1), reflecting the model's ability to accurately detect harmful content without over-penalizing non-hateful posts. After training, we evaluate the model on the held-out test set. This strong baseline achieves a substantial improvement over the majority-class baseline, providing a meaningful benchmark for more advanced models in Milestone 3.

- F1 Score: 0.712