Toxicity & Hate Speech Classification

Leveraging Transformer-Based Models on the MetaHate Dataset

Ashwin Balaji, Ravi Raghavan, Dhruv Verma, Raafae Zaki

Challenge: Nuance in Hate Speech

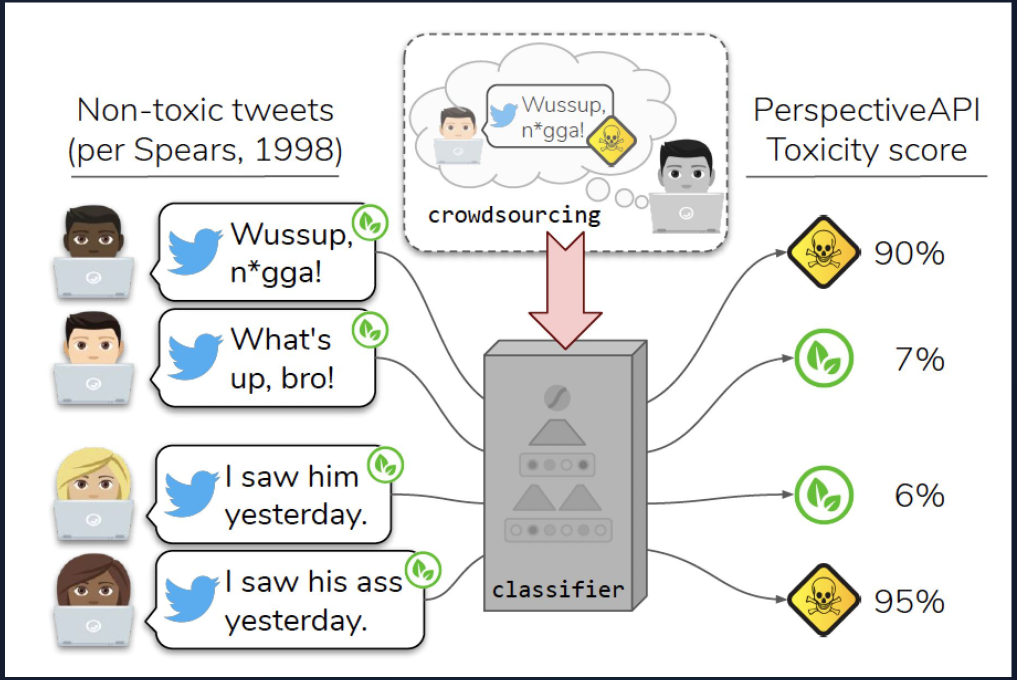
Overt vs. Covert

Detecting hate speech isn't just about keyword matching.

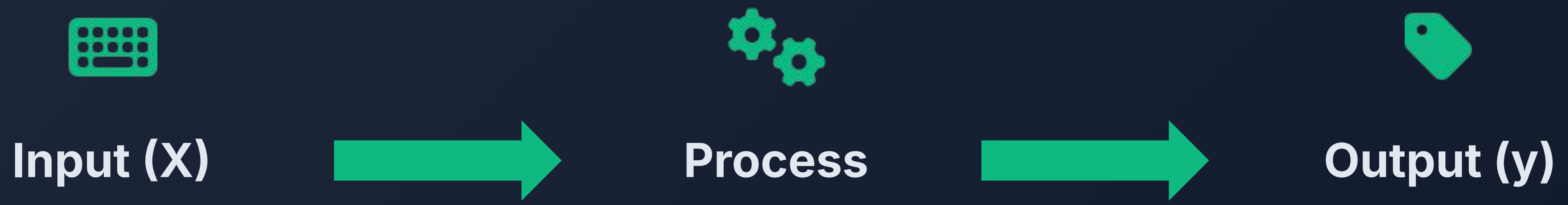
The real challenge lies in context.

> *"I hate [Group X], they are the worst."*
>
> **EASY - directly uses negative keywords**

> *"They should just go back to their country."*
>
> **HARD - implicit/coded language**



Formal Problem Definition



Input (X)

$X = \{W_1, W_2, ..., W_\square\}$

Sequence of text representing a social media post (e.g., tweet, comment)

Process

$f(X) = P(y|X)$

Binary classification model (Transformer) that maps input text to a probability distribution

Output (y)

$y \in \{0: \text{Non-Hate}, 1: \text{Hate}\}$

Binary label indicating the presence of toxicity

Purpose

Massive Scale

Social platforms generate billions of posts daily. Manual moderation is impossible. We need **scalable**, **automated** AI solutions to keep online spaces **safe**.

Real-World Impact

Unchecked hate speech leads to real-world **harassment**, **discrimination**, and even **violence**. We need accurate **censorship** while still preserving **freedom of speech**.

Main Challenge

Building systems that are robust against sarcasm, slang, and evolving cultural contexts

Course Concepts

Transformers

Utilizing the BERT architectures to leverage contextual embeddings that capture word meaning based on surrounding text

Fine-tuning

Re-training a pre-trained language model (BERT) on domain-specific data (MetaHate) for toxicity classification

Handling Imbalance

Using class-weighted cross-entropy loss with F1 to evaluate performance on the imbalanced dataset

Past Approaches

HateBERT

- Fine-tuned BERT on over 1 million Reddit comments from abusive communities
- Domain-specific training significantly boosted F1
- Performance depends on label similarity

Ensemble Methods

- Combined BERT with CNNs for local context
- Combined BERT with LSTMs for sequential context
- Both outperform plain BERT with higher macro-F1 scores
- Computationally expensive

ELECTRA

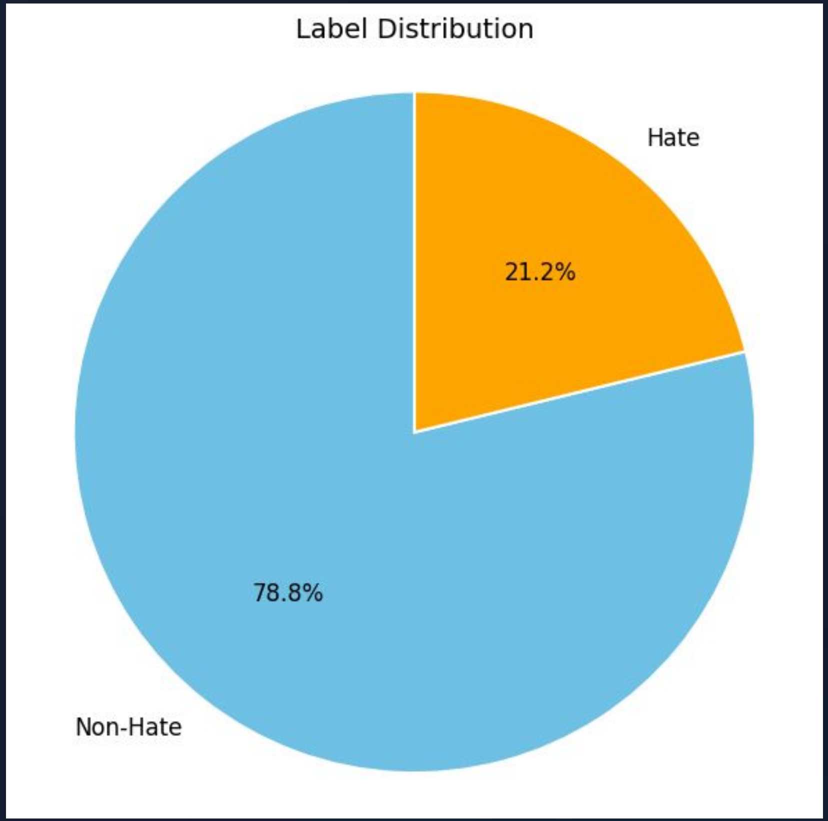
- Achieved SOTA results (~0.89 F1) on MetaHate dataset
- Generator-discriminator framework allowed for implicit hate speech detection
- Still struggles with detecting sarcasm and figurative language

Data: MetaHate Dataset

Unified Benchmark

A meta-collection of **36 different hate speech datasets**, providing a comprehensive view of online toxicity.

- **Total Posts:** ~1.2 Million
- **Format:** TSV (Tab Separated)
- **Challenge:** Significant Class Imbalance



~80% Non-Hate vs. ~20% Hate across Train, Test, and Dev datasets

Evaluation Metric: F1

Accuracy Trap

In an 80/20 dataset, a model that predicts "Non-Hate" for everything achieves 80% accuracy.

Does not actually detect any hate speech

Solution: F1 Score

We use the F1 score for the "Hate" class (1) to balance precision & recall while identifying hate speech specifically

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 ensures that performance evaluation reflects the model’s ability to correctly detect hateful content while minimizing false positives, resulting in meaningful evaluation with an imbalanced dataset

Simple Baseline Performance

Majority Class Classifier

A naive model that simply predicts the most frequent class in the training data for every single input at test time.

Ignores all text features

Since the majority class in the train set was 0 (Non-Hate), this classifier simply predicted 0 for every input in the test set

0.0

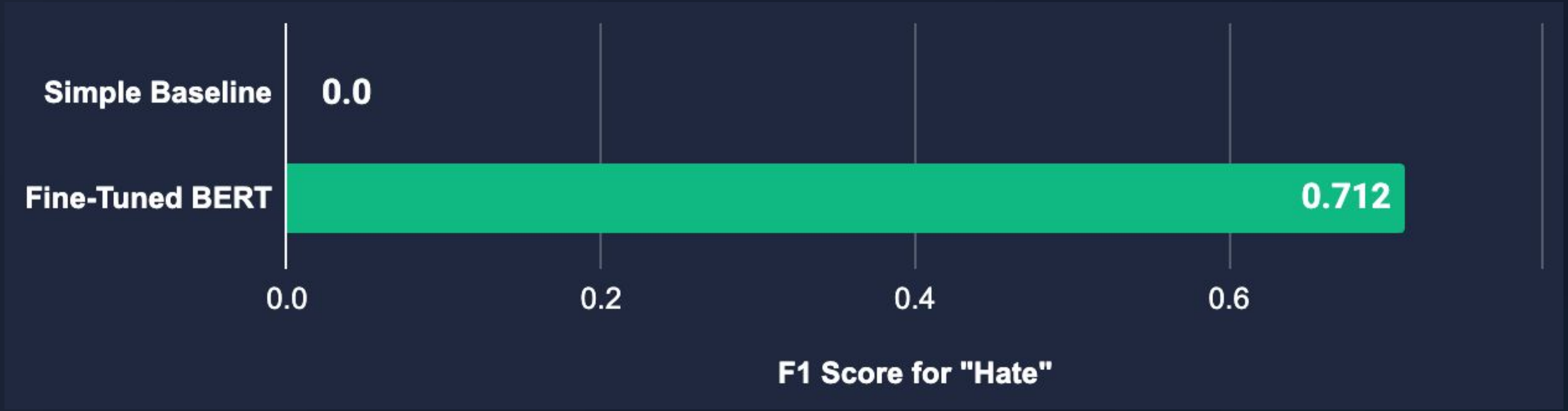
F1 Score for “Hate” (1)

(Accuracy was ~80%, proving it misleading)

Strong Baseline Performance

We fine-tuned **bert-base-uncased** on the MetaHate dataset using

Class-Weighted Cross-Entropy Loss

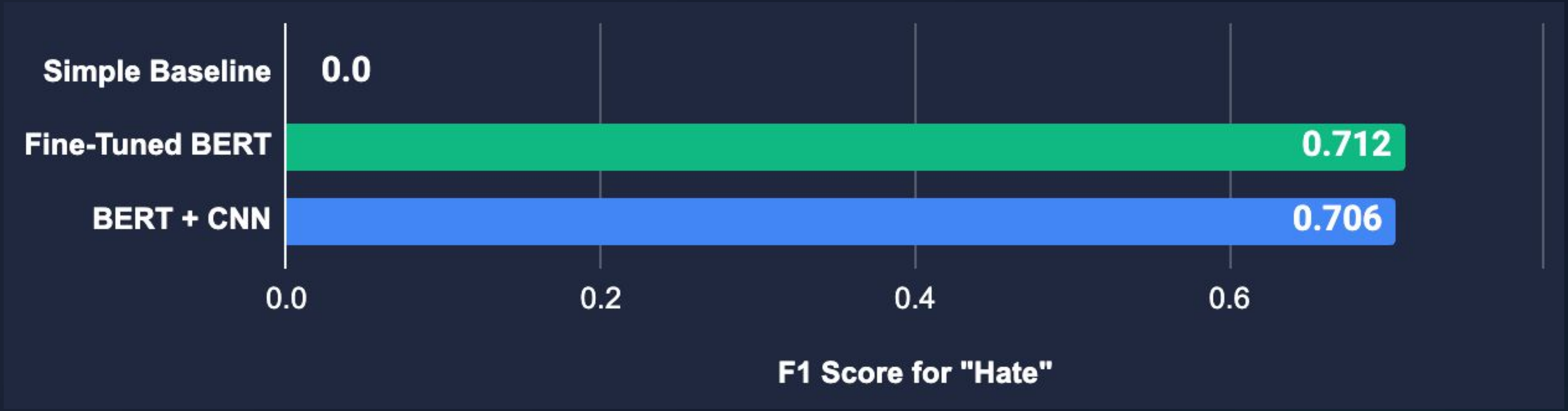


A massive improvement indicating the model is successfully learning linguistic patterns for detecting hateful content

Extension #1

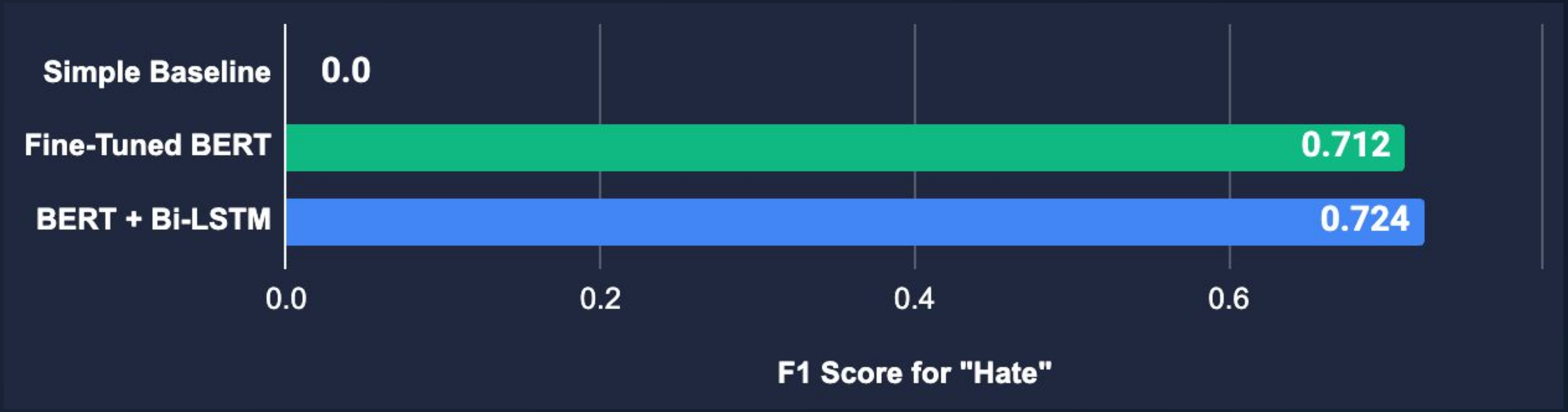
We fine-tuned a **BERT + CNN** hybrid model on the MetaHate dataset using

Class-Weighted Cross-Entropy Loss



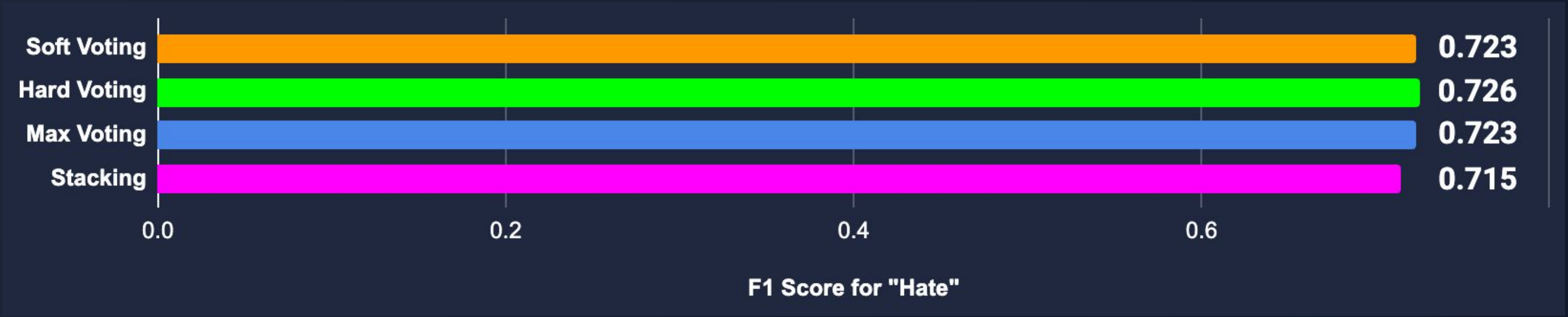
Extension #2

We fine-tuned a **BERT + Bi-LSTM** hybrid model on the MetaHate dataset using **Class-Weighted Cross-Entropy Loss**



Ensemble Approaches

Given the previous extensions, we attempted the following **ensemble** approaches



Error Analysis - Hard Voting Ensemble

False Positives

- Abuse != Hate
  (Eg: “Nobody gives a f*** what you do b****!”)
- Reclaimed or in-group slur usage
  (Eg: “this n***** a clone”)
- Quoted or Reported Hate Detected
  (Eg: “Disabled people are disgusting is a comment that should not be allowed on any social network”)

False Negatives

- Implicit or Ideological Hate
  (Eg: “Just take two tampons and a glass of white wine already”)
- Lexical Evasion via Obfuscation
  (Eg: “Our city is getting flooded with n!****”)

Error Analysis - Extension vs Baseline

Extension Correct | Baseline Wrong

- True Positives Gained

  (True = 1, Ensemble = 1, Baseline = 0)

  - Explicit Slur & Abuse Recognition (Eg: @user c**** c**** w*** w***")
- True Negatives Gained

  (True = 0, Ensemble = 0, Baseline = 1)

  - Social-Media Mentions
  - (Eg: "@user the greatest team...")

Extension Wrong | Baseline Correct

- False Positives Introduced

  (True = 0, Ensemble = 1, Baseline = 0)

  - Sexual Vulgarity (Eg: "...f*** it and get a crispy a** jogging suit")
- False Negatives Introduced

  (True = 1, Ensemble = 0, Baseline = 1)

  - Violence or Death toward protected groups (Eg: "I applaud the m***** of refugees")

Conclusion

Hard Voting Classifier

An ensemble model built upon the previous BERT, BERT + CNN, and BERT + Bi-LSTM models

Combines predictions from multiple models and outputs the class that receives the majority vote.

0.726

F1 Score for “Hate” (1)

THANK YOU