Context-Aware Hate Speech Detection: Traditional Models vs LLMs

Connor McGraw (cmcgraw), Nick Rinaldi (nickrinaldi), and Faiz Hilaly (faiz)

Abstract

1 Introduction

1.1 Project Summary

The internet has become the world's public square, and its scale has outgrown the bandwidth of human moderation. Platforms must sift through billions of comments each day to detect hate speech and harassment without infringing on fundamental rights. To achieve this, companies rely on models to generate confidence scores with respect to the harmfulness of posts. As social platforms increasingly rely on machine learning to make these judgments, the quality of their confidence estimates becomes a matter of civic importance. Using the correct model for this job is essential in shaping how truth and safety are balanced online.

Content moderation systems rely on model confidence to decide when to act. For example, a comment flagged with 90% confidence may be removed completely, while one with 60% confidence will be flagged for human review. The confidence score implies that a score of 0.9 would result in the comment being harmful 9/10 times, but that's not always the case. CNN-based classifiers and LLMs tend to misrepresent their certainty, either overstating or understating it as a function of their bias. Both errors distort decision thresholds and can shift the balance between safety and free expression. This project examines that boundary by looking at whether LLMs or CNNs can better calibrate at the level of confidence real moderation systems use to make decisions.

Our findings can guide how platforms set decision thresholds in automated moderation. Safety teams at companies like Google and OpenAI could use these results to decide what model and methodology ultimately generates the most optimal moderated experience online.

We will use the human-labeled Twitter hate-speech dataset from Kaggle.

1.2 Ethical Considerations

The issue of content moderation has enormous implications, especially as the primary sources of information increasingly shift toward independent contributors on social media platforms. The line between free speech and misinformation is blurry, but it is the job of the models in this study to walk it.

In the case of the CNN we're using, Google's Perspective API has trained their model on millions of comments across diverse sources. They use ten taggers to calibrate their model's output: "if 3 out of 10 raters tagged a comment as toxic, we train the API models to provide a score of 0.3 to this and similar comments" [2]. Given this training methodology, Google's immense resources, and their investment in AI, we have confidence that the model we use will be as unbiased as reasonably possible.

As our LLM, we will be using OpenAI's GPT-5 model, which is the current state of the art. OpenAI has done extensive evaluation around political bias, ultimately finding that "less than 0.01% of all ChatGPT responses show any signs of political bias" [5].

Our aim is to draw a conclusion as to whether state-of-the-art CNNs or LLMs are better for the task of content moderation. While biases can be a factor in each model's performance, that will be factored into our evaluation and will not unknowingly impact the result.

2 Related Work

Davidson et al. [1] introduced the foundational Twitter dataset we use, distinguishing hate speech from offensive language through crowdsourced labels. Their work established classification benchmarks but did not examine model calibration—whether predicted confidence scores reflect true probabilities of correctness.

Commercial systems like Google's Perspective API represent current standards for automated moderation, using CNN and BERT architectures trained on toxic content. Recent work has explored LLMs as alternatives: Guo et al. [3] found GPT-4 achieved competitive accuracy with traditional approaches, while Zahid et al. [6] showed LLMs can capture cultural nuances that simpler models miss. Machlovi et al. [4] advocated for human-first AI moderation with unified benchmarks.

Our work builds upon these efforts by directly comparing the calibration and reliability of CNN-based versus LLM-based moderation systems in realistic decision thresholds.

3 Methodology

3.1 Project Preparation and Prerequisites

We will evaluate the performance of LLMs versus traditional moderation models using the Hate Speech and Offensive Language Twitter dataset, a widely used dataset consisting of approximately 25,000 tweets labeled by human consensus as containing hate speech (0), offensive language (1), or neither/safe (2). Predictions for these tweets will be generated by three systems: Google's Perspective API (a CNN + BERT-based model), and two state-of-the-art LLMs: ChatGPT (GPT-5) and Claude 3.

3.2 Evaluation Process

During evaluation, tweet labels will be hidden from the models, and each system will produce a probability or confidence score for its classification. To enable direct comparison, each model's output will be normalized to the dataset's 0–2 severity scale. The classifications will be compared against human-annotated labels to measure error. We will evaluate both classification performance and calibration quality through measures such as F1 score, ROC AUC, and Expected Calibration Error (ECE), as well as conduct targeted qualitative analysis of specific cases.

Findings will be presented through reliability curves, confusion matrices, and representative example predictions. We expect that LLMs will perform better overall and that their predictions will better reflect the ground truth, given their exposure to broader data and nuanced linguistic context. However, Perspective's simpler architecture may yield more predictable results and fewer outliers.

References

[1] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated Hate Speech Detection and the Problem of Offensive Language". In: *Proceedings of ICWSM*. 2017. URL: `https://ojs.aaai.org/index.php/ICWSM/article/view/14955`.

[2] Google. *About the API Training Data*. `https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US`. Accessed October 2025.

[3] Kun Guo, Anqi Hu, Jiaqi Mu, Zhiyu Shi, Zhiqi Zhao, Nishant Vishwamitra, and Hongxin Hu. *An Investigation of Large Language Models for Real-World Hate Speech Detection*. 2024. arXiv: `2401.03346 [cs.CL]`. URL: `https://arxiv.org/abs/2401.03346v1`.

[4] Noam Machlovi, Mahsa Saleki, Isaac Ababio, and Rayan Amin. *Towards Safer AI Moderation: Evaluating LLM Moderators Through a Unified Benchmark Dataset and Advocating a Human-First Approach*. 2025. arXiv: `2508.07063 [cs.CL]`. URL: `https://arxiv.org/abs/2508.07063`.

[5] OpenAI. *Defining and Evaluating Political Bias in LLMs*. `https://openai.com/index/defining-and-evaluating-political-bias-in-llms/`. Accessed October 2025. 2024.

[6] Abu Hanif Zahid, M. K. Roy, and S. Das. *Evaluation of Hate Speech Detection Using Large Language Models and Geographical Contextualization*. 2025. arXiv: `2502.19612 [cs.CL]`. URL: `https://arxiv.org/abs/2502.19612v1`.