Offensive Language Detection and Bias Analysis

Jonah Smith

JWS130@pitt.edu

Abstract

This paper examines automated offensive language detection and its associated ethical challenges. Using datasets from SemEval-2019 and TwitterAAE, I evaluate an off-the-shelf PerspectiveAPI toxicity classifier alongside a custom logistic regression model. The PerspectiveAPI baseline achieves high precision but exhibits low recall and clear racial dialect bias. After feature and threshold tuning, the logistic regression model surpasses the 70% macro F1 requirement, achieving balanced accuracy and improved sensitivity. The results underscore the persistent trade-off between fairness and detection effectiveness in automated moderation systems.

1 Introduction

Online platforms face mounting pressure to moderate harmful speech while maintaining open expression. Machine-learning classifiers have become essential tools in this space, yet their decisions can reflect and amplify existing social biases. This project explores how model architecture, feature design, and threshold choices affect both predictive performance and fairness. Two systems are compared: (1) the PerspectiveAPI toxicity model, which uses predefined toxicity scores, and (2) a logistic regression classifier trained on SemEval-2019 data. Both are evaluated for classification accuracy, F1 scores, and demographic false-positive disparities to assess technical and ethical trade-offs.

2 Data

The data used include:

- **train.tsv, dev.tsv**: Tweets labeled as `OFF/NOT` from SemEval-2019 Task 6.
- **mini_demographic_dev.tsv**: Tweets labeled by dialect (AA, White, Hispanic, Other), assumed non-offensive for FPR evaluation.
- **test.tsv**: Mixed samples used for final predictions.

Following prior studies (Sap et al.2019; Davidson et al.2017), demographic annotations are used exclusively for fairness assessment. The metrics reported are accuracy, precision, recall, F1 score, and false positive rate (FPR) by group.

3 PerspectiveAPI Baseline

Tweets with *perspective_score* > 0.8 were labeled as offensive. The baseline model achieved an accuracy of 0.76 and a macro F1 of 0.67, but recall for the offensive (OFF) class was only 0.33. This indicates a conservative bias—identifying only explicit toxicity while missing subtler forms of abuse.

Despite its strong precision (0.89), the model's high threshold limits coverage, reflecting a central moderation dilemma: stricter thresholds reduce false positives but allow more harmful content to persist. Threshold tuning can improve recall but risks increasing user over-censorship, illustrating that parameter adjustments have direct social implications.

3.1 Bias Analysis

When applied to the demographic dataset, the model misclassified 19% of African-American English (AAE) tweets as offensive, compared to 7% of White dialect tweets. This mirrors findings from (Sap et al.2019), who show that language models trained on unbalanced web corpora tend to associate AAE syntax and vocabulary with toxicity. Such results emphasize the need for fairness auditing before deployment, as pretrained models may unintentionally penalize minority dialects.

4 Custom Logistic Regression Classifier

To create a transparent and interpretable baseline, I trained a logistic regression classifier using TF-IDF features. Logistic regression was chosen for its simplicity, interpretability, and low variance, allowing precise control over feature weighting and threshold behavior.

After tuning the TF-IDF parameters (sublinear scaling, 20k vocabulary), increasing the weight of the OFF class, and lowering the decision threshold from 0.5 to 0.45, the model achieved 75.3% accuracy and a macro F1 of 0.7098 on **dev.tsv**. The OFF-class F1 improved to 0.60, reflecting higher sensitivity. Although AAE tweets still exhibited a higher FPR (35%) than White tweets (24%), overall disparities were reduced relative to the baseline. These improvements demonstrate that modest feature engineering and threshold calibration can enhance both accuracy and fairness without resorting to complex architectures.

4.1 Comparison and Interpretation

The tuned logistic regression model outperformed PerspectiveAPI in both accuracy and balanced F1, while slightly narrowing demographic bias. However, fairness improvements were partial—AAE tweets continued to experience higher misclassification rates. This outcome aligns with (Davidson et al.2017), who found that lexical models often overfit to slang and profanity tokens common in AAE, conflating informal expression with hostility. Consequently, improvements in recall come at the cost of increased false alarms for marginalized language styles. These findings reaffirm that offensive-language detection is not solely a technical task but also a socio-linguistic challenge that requires equitable model design and representative data.

Beyond quantitative gains, the logistic regression model's interpretability was particularly valuable. Examining TF-IDF feature weights revealed that tokens directly associated with explicit slurs and profanities dominated high positive coefficients, whereas contextual cues such as negations ("not," "never") or quoted speech reduced predicted offensiveness. This transparency allowed identification of linguistic artifacts that would be opaque in transformer-based systems. In contrast, neural architectures like BERT or RoBERTa might learn implicit associations that are difficult to audit. Thus, while the logistic model lacks representational depth, its explainability provides a baseline for fairness auditing and for guiding future bias-mitigation research.

5 Ethical Implications

Defining "offensiveness" is inherently subjective, rooted in cultural and contextual interpretation. Misclassification has unequal consequences: false negatives allow harmful speech to persist, while false positives disproportionately silence underrepresented communities. The disparities observed here illustrate how biased training data and limited demographic representation can compound inequity.

To mitigate these harms, future systems must incorporate diverse data sources, transparent auditing procedures, and human oversight in moderation pipelines. Fully automated censorship risks reinforcing majority language norms, while hybrid systems can balance consistency with contextual understanding. As (Sap et al.2019) argue, achieving fairness in NLP requires iterative evaluation and active collaboration with affected communities rather than post hoc technical fixes.

These findings intersect with broader debates in platform governance. Automated moderation systems often operate at massive scale, where decisions about "acceptable speech" become infrastructural rather than contextual. Biases that silence certain dialects or communities can shape public discourse and erode trust in moderation frameworks. Defining offensiveness, therefore, cannot rest solely on statistical performance; it must involve participatory annotation, continuous feedback loops, and transparency about model limitations. Incorporating community reviewers or appeal mechanisms could provide the sociotechnical balance that purely algorithmic systems currently lack.

6 Conclusion

The PerspectiveAPI baseline demonstrated strong precision but significant bias and low recall, while the tuned logistic regression model achieved 75% accuracy and 0.71 macro F1, satisfying performance targets. Fairness improved modestly but remained imperfect. These results illustrate that careful feature design and threshold tuning can enhance performance, yet systemic bias cannot be eliminated without more representative data and bias-aware training. Future work should explore transformer-based or adversarially debiased models to reconcile performance and equity.

References

[Davidson et al.2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

[Sap et al.2019] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of ACL*.

Appendix: Evaluation Tables

Model	Accuracy	OFF F1	NOT F1	Macro F1
PerspectiveAPI	0.764	0.485	0.847	0.666

Table 1: PerspectiveAPI baseline results on **dev.tsv**.

Demographic	FPR
AA	0.19
Hispanic	0.10
White	0.07
Other	0.01

Table 2: PerspectiveAPI FPR by demographic.

Model	Accuracy	OFF F1	NOT F1	Macro F1
LogReg (tuned)	0.753	0.600	0.820	0.7098

Table 3: Tuned Logistic Regression results on **dev.tsv**.

Demographic	FPR
AA	0.35
Hispanic	0.27
White	0.24
Other	0.02

Table 4: Tuned Logistic Regression FPR by demographic.