Classifier Performance for Twitter Hate Speech Detection

Wanjing Hu

Abstract

This report analyzes two classifiers, a custom NER+BERT pipeline and a fine-tuned RoBERTa model, for Twitter hate speech detection. Both models were trained on 90% of a dataset and evaluated on 10%. While both achieve high in-domain accuracy (96.3% and 98.4%, respectively), analysis reveals critical differences in calibration and generalizability. The NER+BERT model suffers from extreme overconfidence (ECE 0.3795), rendering it unreliable. The RoBERTa model is exceptionally well-calibrated (ECE 0.0152). However, both models fail to generalize to out-of-distribution (OOD) data, with accuracy on the 'Ethos' dataset dropping to 61.5% and 67.9%, respectively.

1 Introduction

Automated hate speech detection is a complex NLP task, challenged by subjectivity, context, and coded language. This report evaluates two Transformer-based models:

- **NER+BERT:** A pipeline that first extracts Named Entities (NER) and then uses BERT for classification.
- **RoBERTa:** A fine-tuned `cardiffnlp/twitter-roberta-base-hate` model.

Both models used weighted-class loss to handle data imbalance during training.

2 Results and Analysis

All models were evaluated on a 10% test split (6,392 samples) for in-domain metrics. The latency is analyzed on A100 GPU.

2.1 Comparative Performance

The comprehensive results across all test scenarios are consolidated in Table 1.

2.2 Error and Calibration Analysis

- **NER+BERT:** This model is **poorly calibrated**. An ECE of 0.3795 is extremely high, indicating chronic overconfidence. Its confidence scores do not reflect the true probability of correctness and are unreliable.
- **RoBERTa:** This model is **exceptionally well-calibrated**. An ECE of 0.0152 is near-perfect, meaning its confidence scores are trustworthy. Error analysis suggests its "most confident errors" are likely mislabeled examples in the ground-truth dataset.

2.3 Generalizability and Robustness

- **Generalizability:** Both models generalize **very poorly**. On the OOD 'Ethos' dataset, accuracy for NER+BERT plummeted to 61.5% and for RoBERTa to 67.9%. This indicates significant overfitting to the source training data.
- **Robustness:** Both models are **highly robust** to simple, syntactic perturbations. The RoBERTa model was unaffected by lowercasing and minimally impacted by typos (0.4% accuracy drop). The NER+BERT model was similarly robust to typos.

3 Discussion

3.1 Effectiveness and Efficiency

- **Effectiveness:** The RoBERTa model is demonstrably more effective. It has higher in-domain accuracy (98.4% vs 96.3%) and F1-score (0.937 vs 0.866). Critically, its strong calibration (ECE 0.0152) makes it a far more reliable and trustworthy model than the overconfident NER+BERT.

Table 1: Comparative Performance Metrics for Hate Speech Classifiers

Model	Dataset	Accuracy	F1 (Macro)	Recall (Macro)	ECE
NER+BERT	In-Domain (Test)	0.9634	0.8664	0.8760	0.3795
NER+BERT	OOD (Ethos)	0.6152	0.5693	0.5822	0.1008
NER+BERT	Robust (Typos)	0.9847	0.9401	0.9267	0.3733
NER+BERT	Robust (NER Ablation)	0.9947	0.9797	0.9768	0.3754
RoBERTa	In-Domain (Test)	0.9840	0.9374	0.9172	0.0152
RoBERTa	OOD (Ethos)	0.6794	0.6686	0.6672	0.3078
RoBERTa	Robust (Lowercase)	0.9840	0.9374	0.9172	0.0152
RoBERTa	Robust (Typos)	0.9797	0.9200	0.8998	0.0189

- **Efficiency:** There is parallelization of the NER pipeline, no comparative metrics on latency on throughput yet due to tie limit.

3.2 Insights from Error Analysis

The error analysis was crucial. For NER+BERT, the high ECE invalidates the model's high accuracy, as its confidence is meaningless. For RoBERTa, the low ECE and analysis of confident errors suggest the model is even more effective than the metrics show, identifying label noise in the dataset. This highlights that calibration (ECE) is a more critical metric than raw accuracy for model trustworthiness.

3.3 Likelihood of Repetition

The high in-domain effectiveness is **highly unlikely** to be repeated on other hate speech datasets. Both models failed severely on the OOD 'Ethos' dataset, demonstrating "brittleness." This failure to generalize is a common finding for models overfitted to a specific data distribution. They are robust to syntactic noise (typos) but not to the semantic and stylistic shifts present in new, real-world data.

4 Conclusion

The fine-tuned RoBERTa model is superior to the NER+BERT pipeline in accuracy, F1-score, and, most importantly, calibration. However, both models are highly overfitted and fail to generalize, limiting their utility on new data without adaptation.