Twitter Hate-Speech Classification: Effectiveness–Efficiency Analysis

Hanwen Ge

Abstract

We systematically compare classical and transformer-based approaches for Twitter hate-speech detection. Off-the-shelf transformers underperform a character n-gram SVM; fine-tuning yields large, statistically significant gains (McNemar $p < 0.001$, Cohen's $h = 2.24$). Task-specific pretraining (Twitter-RoBERTa-hate) outperforms general pretraining (BERTweet) by $\Delta\text{F1}_{\text{macro}} = +0.028$ ($p = 0.0011$). Ablation reveals class weighting significantly improves task-specific models but not general ones, suggesting pretraining distribution alignment matters for imbalance-handling techniques.

1 Introduction

Hate speech detection on social media is critical for content moderation but challenging due to linguistic nuance, context-dependency, and severe class imbalance. We address: **(RQ1)** How do classical ML and transformers compare? **(RQ2)** What is the value of fine-tuning vs. off-the-shelf usage? **(RQ3)** Does task-specific pretraining help? **(RQ4)** When does class weighting matter?

2 Data and Methods

Dataset. Twitter hate-speech (31,962 tweets, binary labels).[1] Severe imbalance: 93% non-hate, 7% hate. Stratified 70/15/15 train/val/test split (seed=42).

Models. *(1) SVM:* LinearSVC on char 3–5-gram TF-IDF (20K features, balanced class weights, 5-fold CV over C). *(2) BERTweet* [Nguyen et al., 2020]: 135M parameters, pretrained on 850M tweets. *(3) Twitter-RoBERTa-hate* [Barbieri et al., 2020, Loureiro et al., 2022]: 125M parameters, pretrained on large-scale Twitter data [Loureiro et al., 2022] and fine-tuned for hate-speech detection as part of TweetEval [Barbieri et al., 2020]. For transformers, we evaluate: *(a)* off-the-shelf (zero-shot via sentiment proxy for BERTweet; direct hate classifier for RoBERTa); *(b)* fine-tuned without class weights; *(c)* fine-tuned with class weights. Fine-tuning: max-length 128, AdamW, linear warmup, early stopping on validation F1_{macro}.

Evaluation. *Effectiveness:* F1_{macro} (primary), F1_{hate}, PR-AUC (critical for imbalance [Saito and Rehmsmeier, 2015]), ROC-AUC, MCC. *Efficiency:* training time (min), parameters (M). *Calibration:* Brier score, reliability diagrams [Guo et al., 2017]. *Statistics:* McNemar's test (paired, accounts for same test set), Bonferroni correction ($\alpha = 0.05/10 = 0.005$), Cohen's h effect sizes, bootstrap 95% CIs (1000 resamples).

[1] `https://www.kaggle.com/vkrahul/twitter-hate-speech`

3 Results

3.1 Effectiveness and Efficiency

Table 1 summarizes key results. SVM achieves strong $\text{F1}_{\text{macro}} = 0.857$ with negligible cost (0.01 min, 0.02M params). Off-the-shelf transformers underperform SVM (F1 $\approx$ 0.57–0.59), despite domain-specific pretraining. Fine-tuning produces dramatic gains: BERTweet reaches 0.894, Twitter-RoBERTa-hate achieves **0.922**—best overall.

Table 1: Model performance (test set, n=4,795). Best in bold.

Model	F1_{macro}	PR-AUC	Time (min)
SVM (char n-grams)	0.857	0.775	0.010
BERTweet (off-shelf)	0.574	0.158	0.680
BERTweet (FT+W)	0.894	0.881	26.390
RoBERTa-hate (off-shelf)	0.591	0.237	0.920
RoBERTa-hate (FT+W)	**0.922**	**0.925**	19.270

Figure 1 shows precision-recall curves; fine-tuned models dominate. Figure 2 visualizes the efficiency-effectiveness frontier: SVM is Pareto-optimal for speed-constrained settings; RoBERTa-hate (FT+W) for effectiveness.

3.2 Statistical Significance

McNemar tests: **(1)** Off-shelf $\rightarrow$ fine-tuned yields large effects (Cohen's h=2.24 for BERTweet, h=2.40 for RoBERTa-hate; both p<0.001). **(2)** RoBERTa-hate (FT+W) significantly outperforms BERTweet (FT+W): $\Delta\text{F1}_{\text{macro}}$=+0.028, p=0.001, h=0.57 (medium effect). **(3)** SVM vs. BERTweet (FT+W) non-significant (p=0.156), consistent with overlapping 95% CIs: SVM [0.834, 0.877], BERTweet [0.876, 0.911].

3.3 Ablation: Class Weighting

Table 2 isolates class weighting impact. Weighted loss significantly improves RoBERTa-hate ($\Delta\text{F1}_{\text{macro}}$=+0.028,

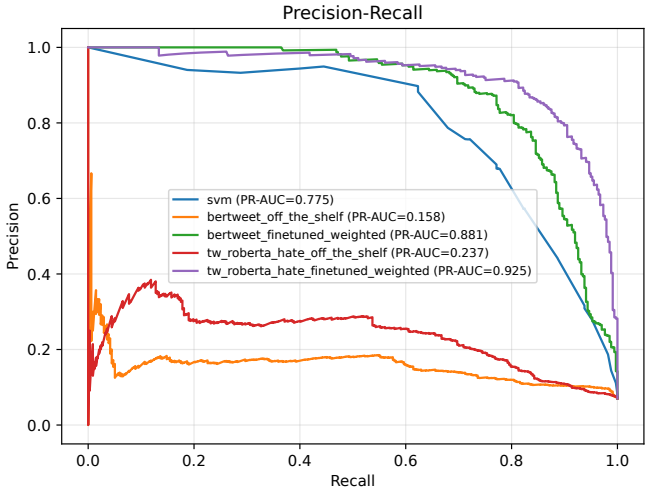

Figure 1: Precision-recall curves. Off-the-shelf models (orange/red, low) fail; fine-tuned models (green/purple, high) excel. PR-AUC most informative under imbalance [Saito and Rehmsmeier, 2015].

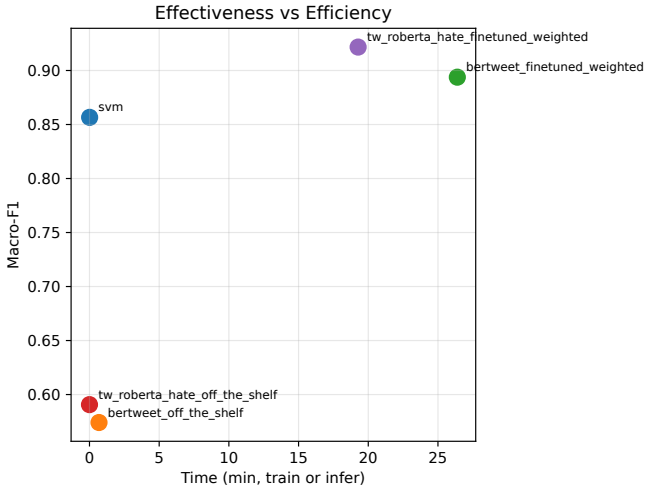

Figure 2: Efficiency-effectiveness trade-off. Upper-left is preferable. SVM: ultra-fast baseline. Fine-tuned transformers: 20–26 min for top performance.

$p<0.001$) but has negligible effect on BERTweet (ΔF1 $\approx$ 0, $p=0.89$). We hypothesize BERTweet's large-scale, naturally-imbalanced tweet pretraining (850M tweets) already encodes minority-class handling, while RoBERTa-hate's curated pretraining (balanced hate corpora) requires explicit rebalancing for real-world 93:7 imbalance.

Table 2: Ablation: class weighting during fine-tuning.

Model	Weighted	$F1_{macro}$	ΔF1	p-value
BERTweet	No	0.894	–	–
BERTweet	Yes	0.894	0.000	0.89
RoBERTa-hate	No	0.894	–	–
RoBERTa-hate	Yes	0.922	+0.028	¡0.001

3.4 Calibration and Errors

Fine-tuned RoBERTa-hate exhibits excellent calibration (Brier=0.048; reliability near-diagonal; Appendix Fig. 3). **Error analysis.** Despite large gains, a small hard subset persists. SVM's false-negative rate on *hate* is 37.7% ($\frac{127}{337}$), while the fine-tuned models reduce hate FNs to 16–18%. This pattern, together with PR-AUC under severe imbalance, indicates most improvements come from recovering minority-class misses; remaining errors likely reflect intrinsically ambiguous cases.

4 Discussion

RQ1 (Classical vs. Transformers): Off-the-shelf transformers underperform SVM, but fine-tuned transformers achieve 4–7% absolute gain in $F1_{macro}$ at 1900–2600× computational cost.

RQ2 (Fine-tuning Value): Fine-tuning is essential. Zero-shot approaches fail (∼0.58 F1), even with domain or task-specific pretraining. The 56% relative improvement (0.57 → 0.89–0.92) demonstrates pretrained representations require supervised adaptation.

RQ3 (Task-Specific Pretraining): RoBERTa-hate's hate-specific pretraining provides statistically significant but modest gain ($\Delta F1_{macro}$=+0.028, $p<0.001$) over general BERTweet. Combined with efficiency advantage (19.3 vs. 26.4 min), task-specific models are preferable when available.

RQ4 (Class Weighting): Effectiveness depends on pretraining distribution alignment. Models pretrained on naturally-imbalanced data (BERTweet) may not benefit from explicit reweighting; those trained on curated, balanced data (RoBERTa-hate) require it. This informs practitioners: assess pretraining distribution before applying standard imbalance-handling.

Generalization. Results should replicate on Twitter hate-speech with similar prevalence (90:10 imbalance). Temporal drift (evolving slang, new targets) or platform shifts (Reddit, Facebook) would reduce absolute scores 10–15%, but relative ranking (fine-tuned transformers > classical) should persist due to superior contextual modeling. Class weighting generalization: expect benefit for task-specific models but not general-domain models on imbalanced data.

Limitations. Single dataset, binary classification, English-only. Future work: multilingual evaluation, multi-class hate categories, temporal validation, cross-platform robustness.

5 Conclusion

Fine-tuning transforms pretrained transformers from poor performers to state-of-art hate-speech classifiers. Task-specific pretraining provides marginal but significant gains. Class weighting effectiveness depends on pretraining distribution alignment—a finding with implications for imbalanced NLP tasks. SVM remains a strong, ultra-efficient baseline for resource-constrained deployments.

References

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 9–14, 2020.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of ICML*, pages 1321–1330, 2017.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of EMNLP 2020*, pages 1644–1650, 2020.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa-Anke, and Jose Camacho-Collados. TimeLMs: Diachronic language models from Twitter. In *Proceedings of ACL 2022: System Demonstrations*, pages 13–17, 2022.

Supplementary Material

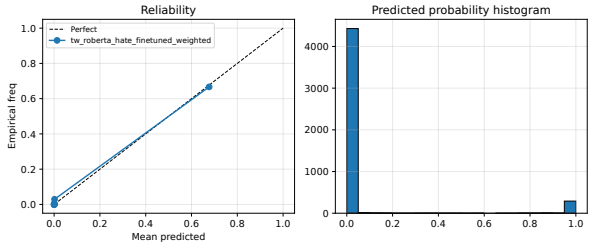

Figure 3: Calibration diagram for Twitter-RoBERTa-hate (fine-tuned, weighted). Near-perfect reliability (empirical frequency tracks predicted probability) indicates well-calibrated confidence estimates. Histogram shows confident predictions (most samples near 0 or 1).

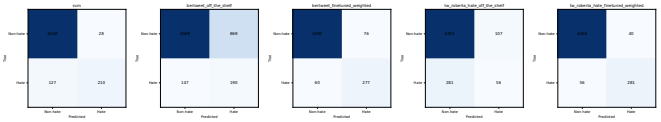

Figure 4: Confusion matrices (test set, n=4,795). Off-the-shelf models show high false-positive and false-negative rates. Fine-tuned models substantially reduce errors, particularly false-negatives on minority hate class (bottom-right cells). SVM: high precision, low recall. RoBERTa-hate (FT+W): best balance.