Hate Speech Classification: Effectiveness–Efficiency Trade-offs on Twitter

Davide Marchi

Master's Degree Student, University of Pisa

Email: d.marchi5@studenti.unipi.it

TABLE I
DATASET OVERVIEW (FROM PROCESSED TRAINING DATA).

n_records	n_classes	avg_tokens	median_tokens
31962	2	12.57352481071272	12.0

TABLE II
LABEL DISTRIBUTION IN THE TRAINING DATA.

label	count	percent
0	29720	92.98542018647143
1	2242	7.014579813528565

***Abstract*—We compare five text classifiers for hate speech detection on Twitter: a regex/keyword baseline, linear SVM with TF–IDF features, MiniLM sentence embeddings with logistic regression, an LSTM sequence model, and a zero-shot classifier based on DistilBERT fine-tuned on MNLI. Effectiveness is measured by macro-F1 and accuracy; efficiency is measured by estimated carbon emissions (kg CO_2e) for training and testing, alongside time. All experiments run on CPU only to ensure comparability, and emissions reflect both runtime and hardware utilization. On the test set, the SVM+TF–IDF model attains the best macro-F1 while remaining among the most emission-efficient models. Pairwise McNemar tests indicate the SVM significantly outperforms the other supervised alternatives. A Pareto analysis of macro-F1 vs test emissions highlights practical trade-offs between effectiveness and environmental cost.**

I. INTRODUCTION

Hate speech detection is commonly framed as a text classification task. In practical settings, model choice should consider not only effectiveness (e.g., macro-F1) but also efficiency and environmental impact. We study this trade-off across five representative approaches, from simple keyword rules to modern pretrained models. The complete code and artifacts are available at: github.com/davide-marchi/hate-speech-model-comparison.

II. DATA

We use the Twitter hate speech dataset (training split) as provided on Kaggle [11]. Table I summarizes corpus-level statistics; class distribution is in Table II. We apply standard text cleaning (URLs, mentions, lowercasing) consistent with reproducible baselines.

III. METHODS

We compare the following models:

- **Regex/keyword**: a lexical baseline built from indicative keywords and patterns; see early rule-based toxic language detection [1].
- **SVM + TF–IDF**: a linear SVM on word/character TF–IDF features, a strong classical baseline for text classification [2], [3].
- **MiniLM + Logistic Regression**: sentence embeddings from Sentence-Transformers `all-MiniLM-L6-v2` [4], [5] followed by logistic regression.
- **LSTM**: a recurrent model over token sequences [6].
- **Zero-shot (DistilBERT-MNLI)**: `typeform/distilbert-base-uncased-mnli` [7] used with an NLI-based zero-shot pipeline [8].

Hyperparameters are selected by cross-validation where applicable, and we report the best validation score together with test metrics (see Table III).

IV. EXPERIMENTAL SETUP

All runs are performed on CPU only for consistency across models and to ensure comparable emissions estimates. We track both *time* and *carbon emissions (kg CO_2e)* during training and testing. Emissions reflect runtime and estimated power draw, which depends on CPU utilization; therefore, we emphasize emissions as an efficiency metric rather than wall-clock time alone [9], [10].

Effectiveness metrics include macro-F1 (primary) and accuracy. For statistical comparisons, we use McNemar's test on paired predictions.

V. RESULTS

A. Effectiveness and Emissions

Table III reports test macro-F1 and accuracy alongside estimated test and train emissions. The SVM+TF–IDF model achieves the highest macro-F1 while keeping emissions low relative to neural alternatives. The keyword baseline is extremely efficient but lags in macro-F1. The zero-shot DistilBERT provides modest macro-F1 and comparatively higher emissions.

TABLE III
EFFECTIVENESS AND EMISSIONS SUMMARY. VALUES PULLED DIRECTLY FROM CSV RESULTS.

model	macro-F1	accuracy	test emi
\csvcoli	0.8642796183779791	0.9648052557484748	8.2121208
\csvcoli	0.7284497168105891	0.8878460816518067	2.2296874
\csvcoli	0.7167586914100906	0.8858126075394963	1.3352817
\csvcoli	0.7048633268008824	0.9357109338338808	7.6879606
\csvcoli	0.5933231889986992	0.8603159705928359	0.00045

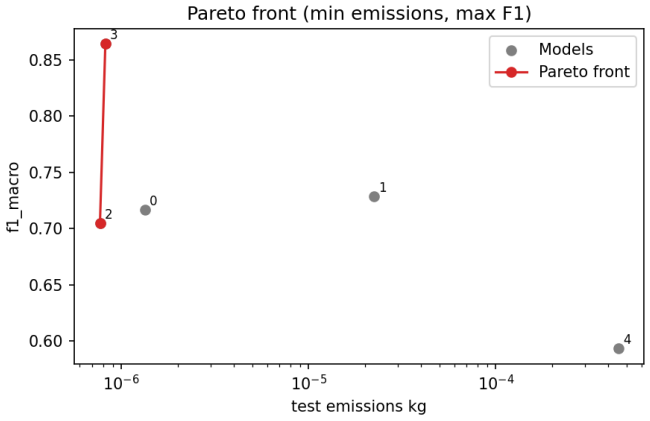

Fig. 1. Macro-F1 vs test emissions with a Pareto frontier.

TABLE IV
MCNEMAR PAIRWISE TESTS (P-VALUE AND SIGNIFICANCE).

model_a	model_b	p_value	significant
\csvcoli	\csvcolii	2.356195889440951e-128	True
\csvcoli	\csvcolii	3.5398989895948825e-92	True
\csvcoli	\csvcolii	1.6874133617076282e-86	True
\csvcoli	\csvcolii	3.593901799380799e-66	True
\csvcoli	\csvcolii	6.524479448423287e-29	True
\csvcoli	\csvcolii	4.319406338856955e-28	True
\csvcoli	\csvcolii	3.7631213497748604e-26	True
\csvcoli	\csvcolii	1.9700388574852112e-07	True
\csvcoli	\csvcolii	1.1495142663259047e-06	True
\csvcoli	\csvcolii	0.6268157481020566	False

Figure 1 shows the Pareto frontier of macro-F1 vs test emissions. We prefer emissions-based comparisons over time-based ones, given emissions encode both runtime and CPU utilization under our CPU-only protocol.

B. Statistical Significance

We apply McNemar's test to paired predictions. Table IV lists pairwise results; SVM+TF–IDF significantly outperforms the neural baselines at $\alpha = 0.05$ on this test set.

C. Error Analysis

We inspect confusion matrices and misclassified examples. Due to space, we include the SVM+TF–IDF confusion matrix in Figure 2; other model-specific reports are similar and are available in the artifacts.

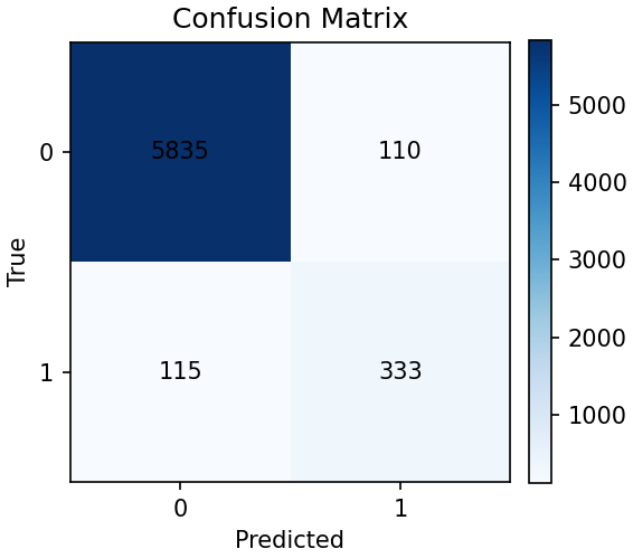

Fig. 2. Confusion matrix for SVM+TF–IDF on the test set.

VI. DISCUSSION

Our results suggest that strong linear baselines remain competitive on short social media text: SVM+TF–IDF provides the best macro-F1 at a favorable emissions footprint. Pretrained embeddings (MiniLM) and LSTM narrow the effectiveness gap but incur higher emissions per test example. Zero-shot classification offers convenience and label flexibility without task-specific training, but its emissions are comparatively higher under CPU-only inference. The keyword baseline is the most efficient but sacrifices recall.

Generalization to other hate speech datasets is plausible for relative rankings between simple linear baselines and larger neural models, though dataset shift (lexical, topical, and label schema) can alter absolute metrics. The emissions trends should persist under CPU-only execution; on GPUs, relative emissions may change depending on batch size and utilization.

VII. LIMITATIONS

Class imbalance and domain-specific language may bias metrics. Emissions are estimates (e.g., based on hardware utilization and regional energy intensity) and should be interpreted as relative comparisons rather than exact measurements. We restrict to small models and CPU to keep runs comparable; larger models or GPUs may change the Pareto frontier.

VIII. CONCLUSION

On this dataset, SVM+TF–IDF is a strong effectiveness–efficiency choice, dominating neural alternatives under CPU-only constraints. Emissions-aware evaluation provides a more informative perspective than time alone because it incorporates utilization. Code and full artifacts are available at github.com/davide-marchi/hate-speech-model-comparison.

REFERENCES

[1] E. Spertus, "Smokey: Automatic Recognition of Hostile Messages," in IAAI, 1997.

[2] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in ECML, 1998.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, 24(5):513–523, 1988.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in EMNLP, 2019.

[5] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," arXiv:2002.10957, 2020.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 9(8):1735–1780, 1997.

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.

[8] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in EMNLP: System Demonstrations, 2020.

[9] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in ACL, 2019.

[10] A. Lacoste, A. Sasha Luccioni, V. Schmidt, and T. Dandres, "Quantifying the Carbon Emissions of Machine Learning," arXiv:1910.09700, 2019.

[11] "Twitter Hate Speech," Kaggle, available at: https://www.kaggle.com/vkrahul/twitter-hate-speech (accessed Nov. 2025).