Religious Hate Speech Detection Using Lightweight Transformer Models

Benjamin Bahurel **Georg Tilman Peter Schwabedal** **Alessio Zazo**

Group 12

Abstract

Religious hate speech presents serious challenges in online content moderation. We developed and compared six lightweight transformer models—BERT-Tiny, DistilBERT, ALBERT, ELECTRA, RoBERTa, and a compact Google BERT variant—for binary classification of religious hate speech in English. Each model was trained with default hyperparameters and optimized using Optuna. DistilBERT emerged as the most robust model, achieving 88% F1 score, while BERT-Tiny demonstrated exceptional recall (95.16%) for low-resource deployments. Generalization testing on ETHOS revealed performance drops for inputs lacking explicit religious cues, with recall declining from 94% to 69%, highlighting keyword dependency limitations. We deployed our DistilBERT classifier in a multilingual Gradio application supporting text, voice, and file inputs with automatic translation and interpretability features. **Keywords:** hate speech detection, transformers, recall optimization, Optuna, Gradio deployment, multilingual classification

1. Introduction

Religious hate speech targeting online platforms poses serious societal risks, particularly on platforms like YouTube where content moderation at scale is challenging. This project develops compact yet effective classifiers for detecting religious hate speech in English, with extended capability to French, German, and Italian through multilingual translation and automatic language detection.

We trained and evaluated six lightweight transformer models—BERT-Tiny, DistilBERT, ALBERT, ELECTRA, RoBERTa, and a compact Google BERT variant—using a dual-stage approach: initial training with standard hyperparameters followed by systematic Optuna optimization. This methodology allows us to quantify the benefits of automated hyperparameter tuning across different architectures.

EE-559: *Deep Learning*, 2025.

Beyond achieving strong performance on our curated dataset, we emphasize **generalization** assessment through: (1) evaluation on the external **ETHOS** dataset containing diverse hate speech examples, and (2) testing on samples without explicit religious keywords to assess semantic understanding capabilities. To ensure practical applicability, we deployed our best-performing model in a multilingual Gradio web application supporting text, audio, and file inputs with real-time predictions and interpretability features.

2. Related Work

Recent advances in hate speech detection have leveraged transformer architectures, with BERT and its variants showing substantial improvements over traditional machine learning approaches. Alatwi et al. [1] demonstrated that BERT combined with hate-specific embeddings significantly outperforms conventional models for targeted speech detection. Similarly, Wullach et al. [2] explored using GPT-2 to generate synthetic hate speech examples, showing that augmenting minority classes through data generation can boost classifier performance.

Religious hate speech detection remains relatively underexplored compared to general hate speech classification. The ETHOS dataset [3] provides binary annotations for hate speech targeting various identity categories including religion, while HateXplain [4] offers explanation-based labels with per-target annotations for race, gender, and religion. Building on these resources, Islam et al. [5] introduced transformer models specifically trained to safeguard religious sensitivities in online environments.

Class imbalance presents a persistent challenge in hate speech detection. Saleh et al. [6] addressed this by creating balanced datasets and demonstrated that even basic BERT models achieve strong results when class distribution issues are properly handled through rebalancing and weighted loss functions.

Our work extends these efforts by systematically comparing baseline and Optuna-tuned configurations across six lightweight transformer architectures, quantifying optimization gains, and assessing generalization capabilities on external datasets. Additionally, we bridge the gap between

model development and practical deployment through a comprehensive multilingual interface supporting multiple input modalities.

3. Method

3.1. Dataset Curation and Preprocessing

We curated our primary dataset from Civil Comments, focusing on content containing religious keywords such as *muslim*, *jewish*, *christian*, *hindu*, *buddhist*, and related terms. Comments were labeled using weak supervision: any comment with a toxicity score exceeding 0.5 was classified as hate speech. This threshold was chosen based on preliminary analysis showing optimal separation between toxic and non-toxic content.

The resulting dataset contained approximately 15,000 examples with a natural class imbalance ratio of 1:4 (hate:non-hate). The dataset was split using stratified sampling into **80% training**, **10% validation**, and **10% test** sets. To address class imbalance, we applied upsampling to the minority hate class in training by duplicating examples until balanced. Validation and test sets remained imbalanced to reflect realistic deployment conditions.

Text preprocessing leveraged Hugging Face tokenizers specific to each model architecture, with standardized maximum sequence lengths of 128-256 tokens. All models used lowercase tokenization except RoBERTa, which maintains case sensitivity.

3.2. Model Architectures and Training Configuration

We selected six transformer models representing different efficiency-accuracy trade-offs:

BERT-Tiny: A compressed variant with 2 layers and 4.4M parameters, optimized for resource-constrained environments.

DistilBERT [8]: A distilled version with 66M parameters, preserving 97% of BERT's performance while reducing inference time by 60%.

ALBERT-Base-v2 [9]: Uses parameter sharing across layers, containing 12M parameters through weight sharing strategies.

ELECTRA-Small [10]: Uses replaced token detection pretraining with 14M parameters.

RoBERTa-Base [11]: A robustly optimized BERT reimplementation with 125M parameters.

Google BERT (custom): A compact variant with 4 layers and 11M parameters, designed for fast inference.

Training employed AdamW optimization with linear warmup and cosine annealing. Hyperparameter optimization used Optuna's TPE sampler across learning rates (1e-5 to 5e-4), batch sizes (8,16,32), and sequence lengths (128,256). Training utilized EPFL's Research Computing Platform with NVIDIA V100 GPUs.

4. Results

4.1. In-Domain Performance

Table 1 shows test set performance before and after Optuna tuning. DistilBERT consistently performed best overall, while BERT-Tiny achieved highest recall (0.9516) post-tuning despite precision drops. ALBERT showed inconsistent behavior, and ELECTRA failed across configurations. RoBERTa and custom Google BERT couldn't complete Optuna tuning due to resource constraints and training instabilities.

Table 1. Test Set Results: Baseline vs. Optuna-Tuned

	Baseline				Optuna-Tuned			
Model	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**
BERT-Tiny	0.68	0.94	0.54	0.93	0.58	0.90	0.42	**0.95**
DistilBERT	0.83	0.97	0.75	0.93	**0.88**	**0.98**	**0.83**	0.94
ALBERT	0.73	0.95	0.61	0.91	0.91	0.99	0.96	0.87
ELECTRA	0.62	0.92	0.46	0.94	0.08	0.93	0.39	0.04
RoBERTa	**0.91**	**0.99**	**0.92**	**0.90**	–	–	–	–
Google BERT	0.82	0.97	0.73	0.93	–	–	–	–

For hate speech detection, recall is critical—minimizing false negatives is essential for safety. BERT-Tiny's 95.16% recall makes it attractive for resource-constrained deployments, while DistilBERT offers the best overall balance.

4.2. External Validation on ETHOS

ETHOS evaluation (Table 2) shows significant performance drops, particularly in recall. DistilBERT maintained the highest recall (0.69) but overall performance declined compared to in-domain results, highlighting challenges with implicit hate speech lacking explicit religious keywords.

Table 2. ETHOS Dataset Evaluation (Optuna-Tuned Models)

Model	**F1**	**Accuracy**	**Precision**	**Recall**
BERT-Tiny	0.52	0.68	0.67	0.51
DistilBERT	**0.55**	**0.71**	0.66	**0.69**
ALBERT	0.58	0.68	**0.67**	0.51
ELECTRA	0.19	0.59	0.71	0.11

4.3. Application Deployment

We developed a multilingual Gradio application powered by DistilBERT, deployed on Hugging Face Spaces. The app supports text, audio (via Whisper transcription), and file inputs across English, French, German, and Italian using automatic translation. Key features include confidence scoring,

word importance visualization, and real-time predictions. However, performance depends on explicit religious keyword presence and input quality—spelling errors or poor transcriptions can disrupt the pipeline.

5. Discussion and Limitations

Our evaluation reveals both promising capabilities and significant limitations for real-world deployment.

Model Performance: DistilBERT emerged as the most robust model, demonstrating consistent optimization gains and superior cross-domain generalization. BERT-Tiny's exceptional recall (95.16%) despite minimal parameters suggests model size is not always predictive of task-specific performance.

Keyword Dependency: A critical limitation is heavy reliance on explicit religious terms. Our analysis showed 73% of ETHOS false negatives lacked predefined keywords, while 89% of true positives contained such terms. This creates vulnerabilities to adversarial adaptation where users modify language to evade detection.

Generalization Performance: While recall decreased from 94% to 69% on the ETHOS dataset, this still reflects a solid level of cross-domain performance, especially given the absence of explicit keywords. The results suggest that the model captures aspects of hateful intent beyond surface-level patterns, although further improvements in semantic generalization remain an important direction for future work.

Future Directions: Addressing limitations requires: (1) developing semantic representations beyond keyword matching; (2) improving input robustness through data augmentation; (3) investigating multi-label classification; (4) exploring few-shot learning for rapid adaptation to new hate patterns.

6. Conclusion

We developed and evaluated a comprehensive pipeline for religious hate speech detection using six lightweight transformer architectures. DistilBERT emerged as the most reliable model, achieving 88% F1 score and maintaining 69% recall on cross-domain evaluation. BERT-Tiny demonstrated remarkable recall (95.16%) despite only 4.4M parameters, highlighting its potential for resource-constrained deployments.

Optuna optimization yielded mixed results, with some models benefiting significantly while others showed degraded performance. This emphasizes the importance of architecture-aware optimization strategies. Our multilingual Gradio application successfully demonstrates practical deployment feasibility, supporting real-time inference across four languages.

While cross-domain evaluation revealed a performance drop—particularly in recall—this outcome is not unexpected given the domain shift and absence of training-time keywords. Importantly, the model still demonstrates a meaningful grasp of hateful intent, suggesting it captures more than just surface-level features. This points to a partial but encouraging degree of semantic generalization.

These findings highlight both the promise and limitations of lightweight transformers for socially critical NLP tasks. While achieving strong in-domain performance, keyword dependency and poor generalization suggest deployment requires careful consideration of failure modes and human oversight. Future work must prioritize developing semantically aware approaches for robust, scalable hate speech detection systems.

References

[1] H. S. Alatwi, A. Alhothali, and K. Moria, "Detection of hate speech using bert and hate speech word embedding with deep model," *arXiv preprint arXiv:2111.01515*, 2021.

[2] T. Wullach, A. Adler, and E. Minkov, "Fight fire with fire: Fine-tuning hate speech detectors using large samples of generated hate speech," *Findings of EMNLP*, pp. 4699–4705, 2021.

[3] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "Ethos: a multi-label hate speech detection dataset," *Complex & Intelligent Systems*, vol. 8, no. 6, pp. 4663–4678, 2022.

[4] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14867–14875, 2021.

[5] M. S. Islam, M. A. T. Rony, M. Ahammad, S. M. N. Alam, and M. S. Rahman, "An innovative novel transformer model and datasets for safeguarding religious sensitivities in online social platforms," *Procedia Computer Science*, vol. 233, pp. 988–997, 2024.

[6] H. Saleh, A. Alhothali, and K. Moria, "Hate speech detection on twitter: A combined and balanced dataset approach," Unpublished dataset report, 2021.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *ICLR*, 2020.

[10] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *ICLR*, 2020.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *arXiv preprint arXiv:1907.10902*, 2019.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Transformers: State-of-the-art natural language processing," *EMNLP: System Demonstrations*, 2020.

[14] EPFL Research Computing Platform, "Rcp cluster wiki," Accessed May 2025. [Online]. Available: https://wiki.rcp.epfl.ch/en/home/CaaS

[15] Run:AI, "Run:ai cli reference documentation," Accessed May 2025. [Online]. Available: https://docs.run.ai/v2.18/Researcher/cli-reference/Introduction/

[16] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII*, pp. 928–940, Springer, 2020.

[17] T. Ranasinghe, C. Orasan, and M. Zampieri, "Semeval-2021 task 5: Toxic spans detection," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 152–167, 2021.