Reading Between the Lines: Fine-Tuning Language Models for Context-Aware Dog Whistle Detection

Lino Zurmühl

Hertie School, Data Science Lab
Supervisor: Prof. Simon Munzert, PhD

Introduction

Dog whistles are coded language used to signal harmful messages to specific groups while appearing innocuous to others. This phenomenon has become increasingly prevalent online as "algospeak" to evade content moderation.

Why Dog Whistles Matter:

- They enable the spread of hate speech while avoiding algorithmic detection
- They contribute to political polarization
- They present unique challenges for content moderation

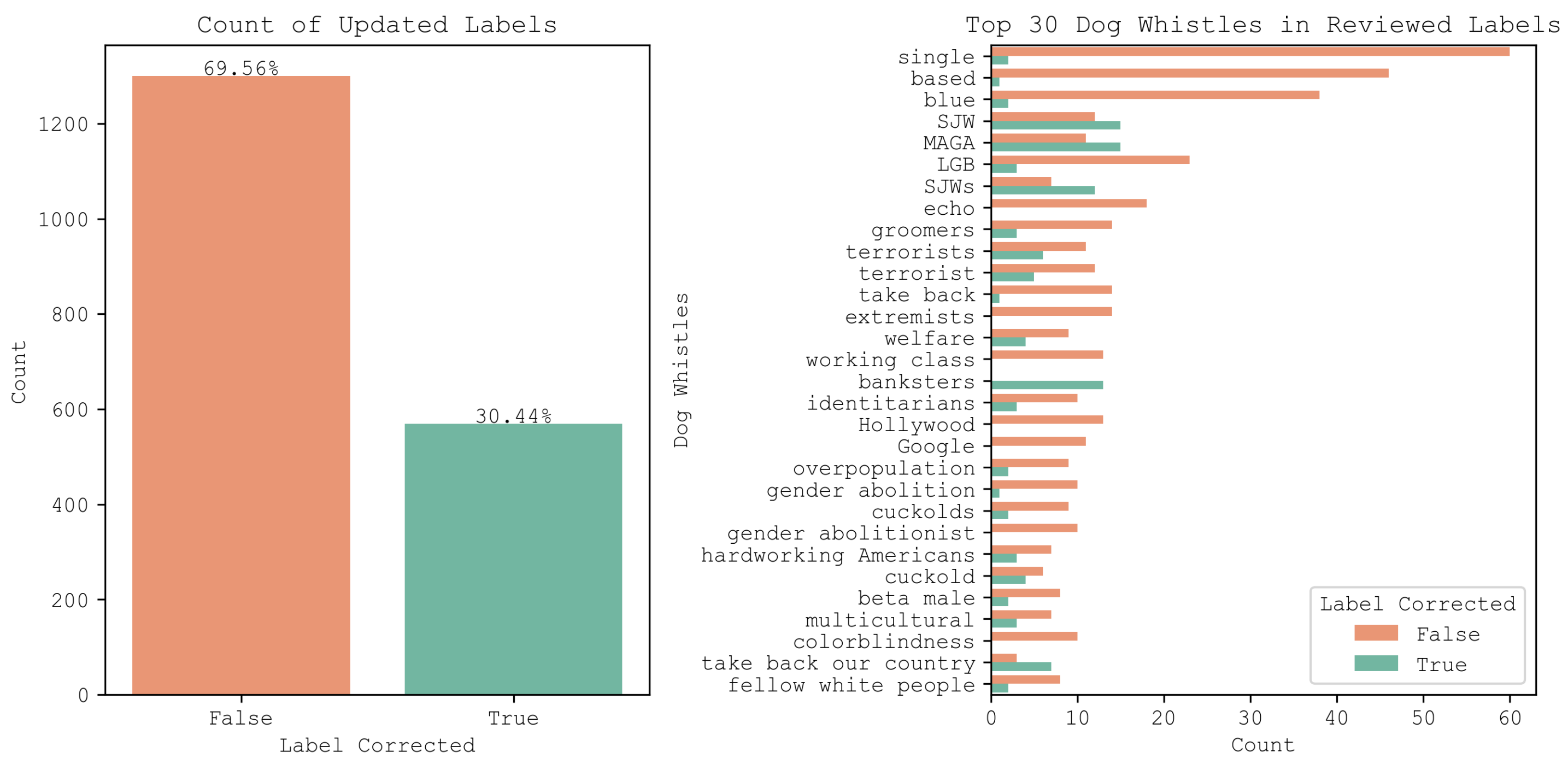
Research Question:
How effective are fine-tuned language models in accurately classifying implicit hate speech characterized as right-wing dog whistles in Reddit posts?

Henderson & McCready (2018) highlight four theoretical perspectives on dog whistles:

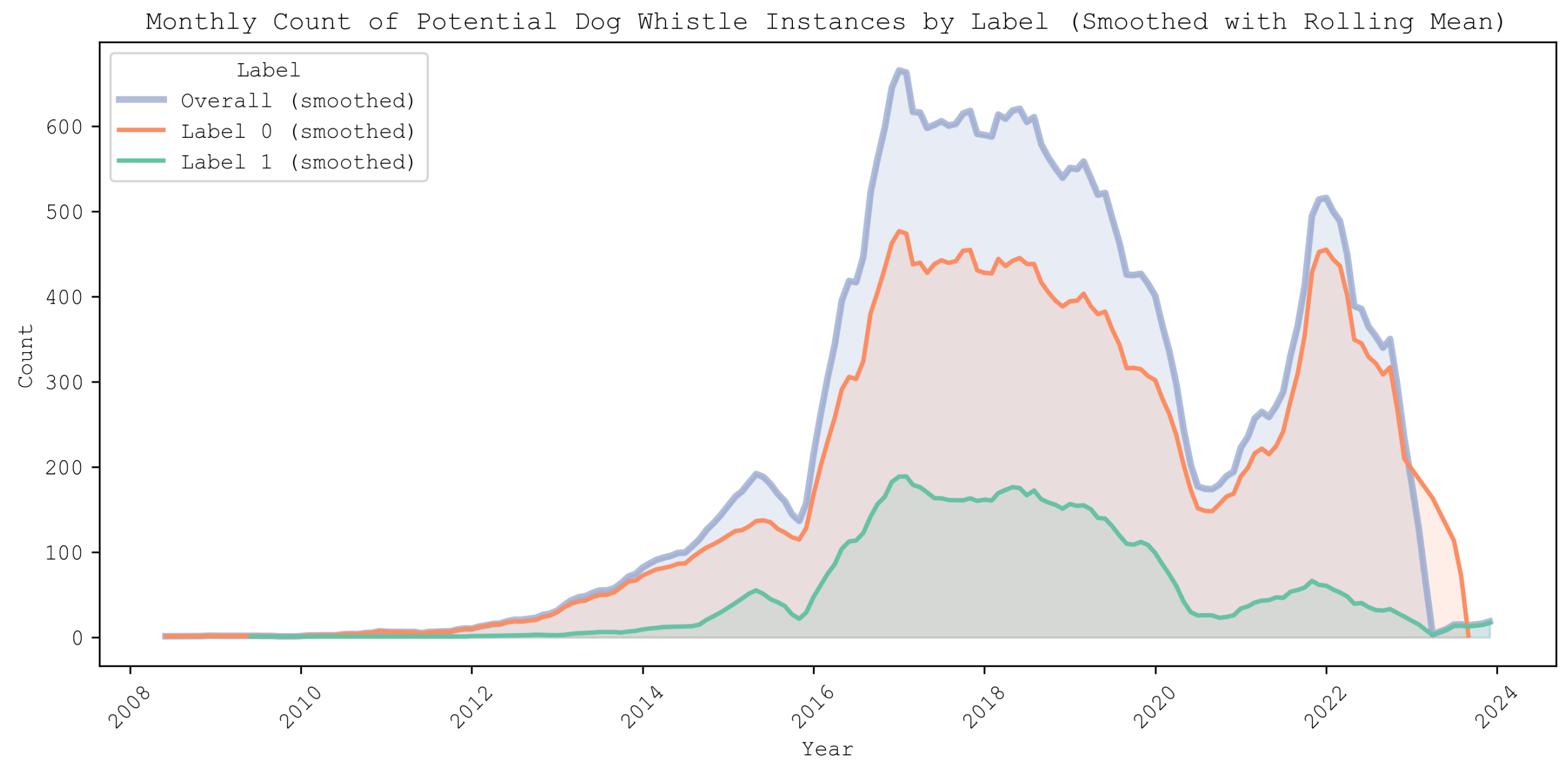
- **Conventional Implicature:** Dog whistles convey hidden meanings encoded within language
- **Inferentialist:** Dog whistles activate listeners' pre-existing beliefs
- **Game-Theoretic:** Dog whistles maintain plausible deniability
- **Mixed View:** Combines inference and strategic design

Methodology

Data: Silent Signals dataset (Kruk et al., 2024):
Data consisting of 10,021 dog whistle texts vs. 32,302 non-dog whistle texts. They were cleaned using Confident Learning (Northcutt et al., 2021)



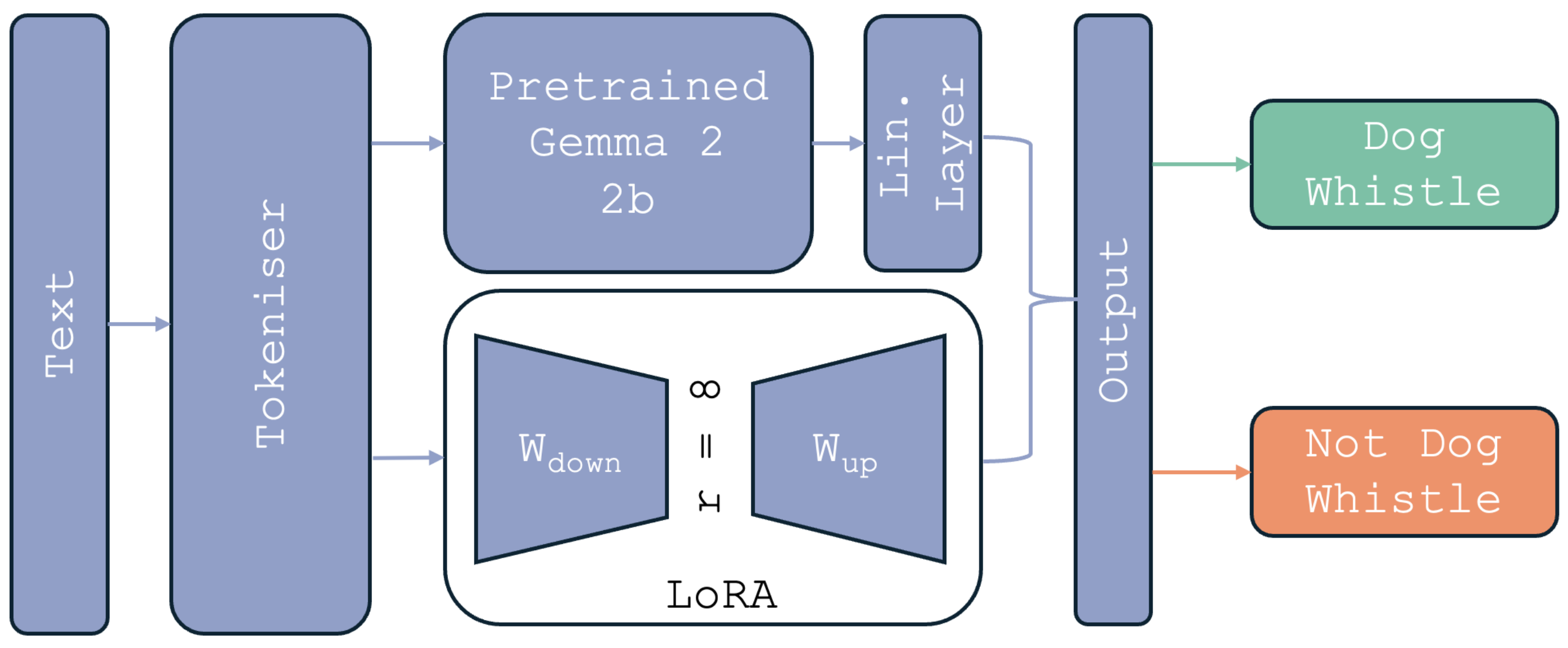
Distribution of Dog Whistle Labels and Manual Corrections



Number of Reddit posts over time by label with moving average of a six month span.

Models:

- **RoBERTa** (355M parameters): Full fine-tuning
- **Gemma 2** (2B parameters): Low-Rank Adaptation (LoRA)



Text

Tokeniser

Pretrained Gemma 2 2b

Lin. Layer

W_{down} $r = 8$ W_{up}

LoRA

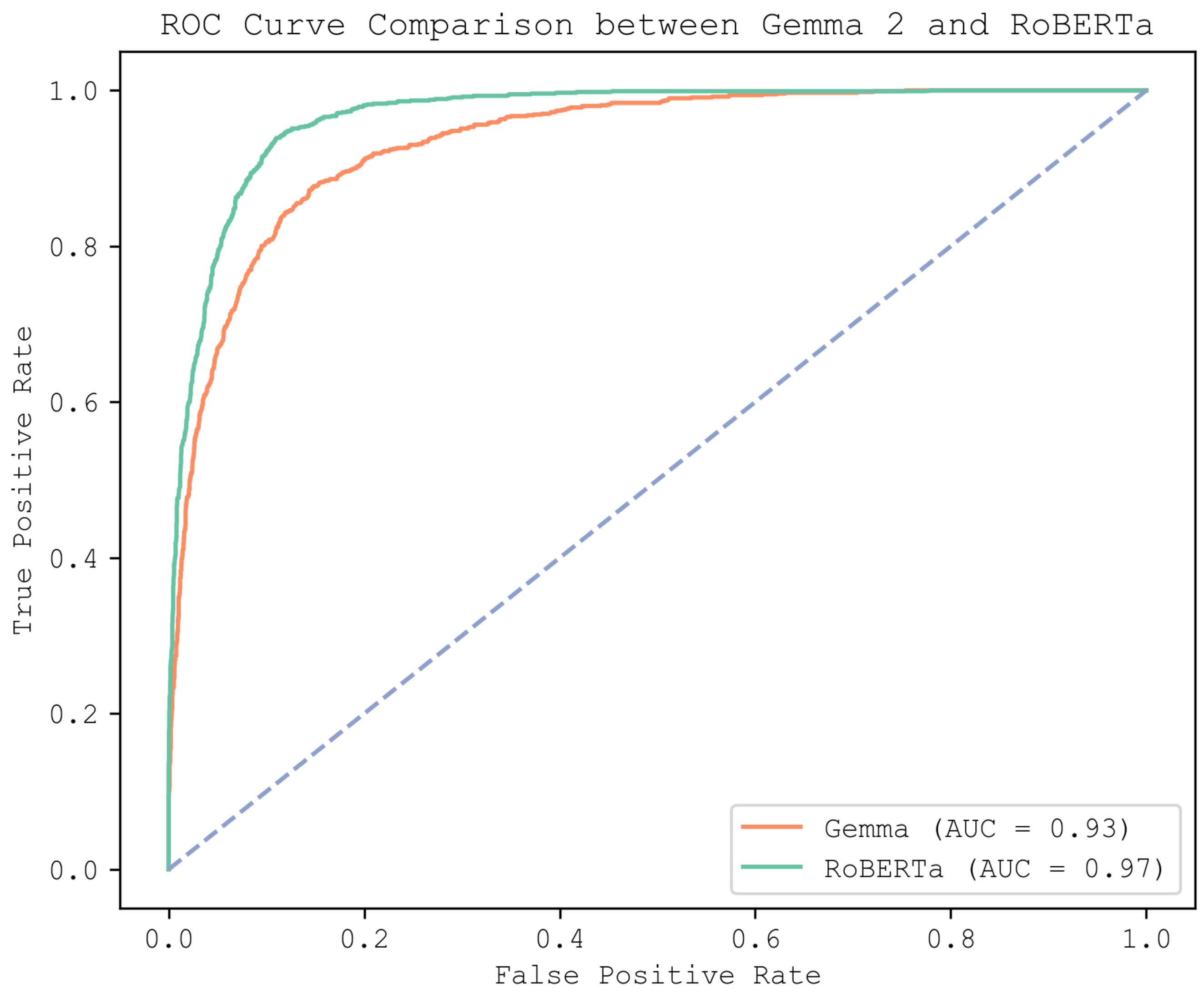
Output

Dog Whistle

Not Dog Whistle

Model Architecture: Fine-tuning using Pretrained Gemma 2 with LoRA

Results



ROC Curve Comparison between Gemma 2 and RoBERTa

In-Distribution:

- RoBERTa: 90% accuracy, F1 = 0.82
- Gemma 2: 88% accuracy, F1 = 0.76

Out-of-Distribution:

- RoBERTa: 46% accuracy, F1 = 0.04
- Gemma 2: 64% accuracy, F1 = 0.51
- Human baseline: 67% accuracy
- GPT-4: 87% accuracy

While RoBERTa shows strong in-distribution performance but struggles with generalization, Gemma 2 demonstrates better cross-domain robustness, though both models underperform compared to humans and GPT-4, partly due to test data mislabeling.

Key Findings & Future Work

Key Findings:

- RoBERTa performs better in-distribution but struggles to generalize
- Gemma 2 shows more robust performance across domains
- Both models underperform compared to human annotators and proprietary systems
- Data quality significantly impacts model effectiveness
- Temporal concentration of data (2016-2020) introduces bias

Future Work:

- More diverse, manually-verified datasets
- Multi-lingual and cross-cultural dog whistle detection
- Real-time adaptation to emerging coded language
- Exploration of other model architectures like modernBERT or Gemma 3

References

- Henderson, R., & McCready, E. (2018). How Dogwhistles Work. In S. Arai et al. (Eds.), New Frontiers in Artificial Intelligence (pp. 231-240). Springer.
- Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.
- Kruk, J., et al. (2024). Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles. arXiv:2406.06840.
- Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.
- Magu, R., et al. (2017). Detecting the Hate Code on Social Media. arXiv:1703.05443.
- Mendelsohn, J., et al. (2023). From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. arXiv:2305.17174.
- Northcutt, C. G., et al. (2021). Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv:2103.14749.
- Team Gemma et al. (2024). Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.
- Witten, K. (2023). The Definition and Typological Model of a Dogwhistle. Manuscrito, 46.