Detecting Emerging Dog Whistles

Madeleine Hueber **Chiara Evangelisti** **Aurélien Genin**

Group 16

Abstract

This project explores the use of language models to detect coded hate speech, known as *dog whistles*, which are often difficult to identify using traditional methods. Using the Silent Signals dataset—comprising labeled political and online text—we fine-tuned RoBERTa with masked language modeling (MLM) oand trained two downstream classifiers for detecting dog whistle usage and identifying the targeted group. The models showed strong performance across tasks, including when tested on previously unseen dog whistles. While results are promising, limitations remain due to the structure of the dataset and fixed class definitions. This work contributes to the growing field of automated hate speech detection, to foster a safer online environment.

Keywords: dog whistle, coded words, in-group

1. Introduction

Dog whistles are a type of coded language – words or expressions – that carry hidden meaning recognizable only inside certain groups, on top of their common use. For example, the number '88' is sometimes used in neo-Nazi circles to represent 'Heil Hitler', since 'H' is the eighth letter of the alphabet. Detecting dog whistles is difficult, even for humans, because their meanings depend heavily on context, and they often change over time.

As new dog whistles appear to replace old ones, it becomes important to develop tools that can help identify them, especially in online settings. In this project, we look at how language models can be fine-tuned to detect both known and emerging dog whistles, and to identify which groups are being targeted. Our aim is to explore whether this kind of approach can support better detection of subtle or evolving forms of coded hate speech.

2. Related Work

Dog whistles notoriously evade traditional hate speech detection because of their implicit and coded nature. Consequently, only a few previous works address the detection and classification of dog whistles.

To support research in this area, some common datasets have been created. The Allen AI Glossary of Dog Whistles [1] gathers 340 english dog whistles along with their definition and targeted in-group. From this glossary, the Silent Signals dataset [2] has been created to give 16k labeled examples of dog whistles, from formal and informal context.

In parallel, some works focused on the binary classification of words to detect whether they are used as a dog whistle or not [3]. While these models achieve good performance, they are limited to a fixed list of coded words and can be quickly out-dated.

Previous works have attempted to detect *emerging* dog whistles [4], but these have typically focused on a narrow set of targeted groups and have yielded limited results.

3. Method

The general pipeline used in this work is summarized in Fig.1. The different stages of the methodology adopted are described in the following sections.

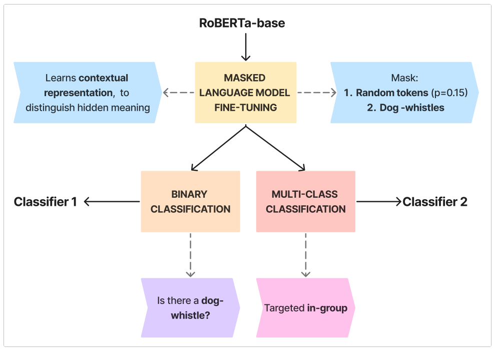

RoBERTa-base
Learns contextual representation, to distinguish hidden meaning
MASKED LANGUAGE MODEL FINE-TUNING
Mask:
1. Random tokens (p=0.15)
2. Dog -whistles
Classifier 1
BINARY CLASSIFICATION
MULTI-CLASS CLASSIFICATION
Classifier 2
Is there a dog-whistle?
Targeted in-group

Figure 1. Schematic representation of the training methodology

3.1. Dataset preparation

Our work used the Silent Signals dataset [2], created from Reddit posts (informal) and US Congress speeches (formal). From a subsample of human-labeled dog whistles, GPT prompts were crafted to label high-fidelity dog whistles in the larger data set. The quality was ensured by the previous researchers by requiring a common positive detection of three different prompts.

From this dataset, we derived our own for the training of our models. We balanced it to have as many dog whistles as non dog whistles, through under-sampling. We also balanced

EE-559: *Deep Learning*, 2025.

the distribution of targeted in-groups to avoid some classes being under-represented. The dataset was finally cleaned of emoji-containing messages, and randomly split into 90% training and 10% validation sets.

Finally, to evaluate the detection of emerging dog whistles, we randomly selected 10% of the dog whistles present in the dataset, and removed all messages containing them from the training set. These messages were used to create a dedicated test set for assessing generalization to unseen expressions.

3.2. Masked language modeling (MLM)

To enhance the model's ability to capture the subtle contextual signals that characterize dog whistle expressions, we adopted a Masked Language Modeling (MLM) objective as an intermediate fine-tuning step. Starting from a pre-trained transformer-based language model (such as BERT or RoBERTa , see Section 4), we further trained the model on our dog whistle dataset using MLM. In this setup, two types of tokens were masked: random tokens across the sentence (with probability 0.15) and the known dog whistle tokens present in the input. The model was then trained to predict the original masked tokens based on their surrounding context.

By learning from patterns, the model is expected to develop context-sensitive representations that can reveal potential hidden meanings embedded in everyday language.

3.3. Classifiers

Following MLM fine-tuning on our dataset, we trained two separate classifiers to detect and characterize dog whistle content.

- **Classifier 1 - Dog whistle detection** : The first classifier was designed as a binary classifier to distinguish between sentences containing or not a dog whistle. It was initialized with the weights obtained from the MLM fine-tuning step and trained for three epochs on our labeled dataset.

- **Classifier 2 - In-group identification** : The second classifier operates only on instances identified as containing a dog whistle by the first model. It performs multi-class classification to predict the specific in-group targeted by the input. Training was performed over five epochs using a labeled dataset where each sample was annotated with its corresponding dog whistle category (e.g., racial, antisemitic, homophobic, etc). This approach enables a more granular understanding of the communicative intent behind each instance.

Together, these classifiers form a two-stage analytic framework: the first stage flags potential dog whistle content, and the second disambiguates its ideological or social target. This architecture facilitates both detection and interpretation of emerging dog whistle language in natural text.

4. Validation

4.1. Model selection

To determine the most suitable transformer architecture for our task, we conducted a series of experiments evaluating three widely-used pre-trained models : BERT-base-uncased [5], RoBERTa-base, and RoBERTa-large [6]. Each model was fine-tuned on our dataset using the same MLM procedure, and subsequently trained on both classification tasks introduced in Section 3. To ensure fair comparison, all models were trained using identical hyperparameter settings.

Tables 1 and 2 report the results on the test set for Classifier 1 (dog whistle detection) and Classifier 2 (in-group identification), respectively. Results are reported over five runs with different random seeds.

Model	Precision	F1-score
BERT-base-uncased	0.892± 0.003	0.892± 0.003
RoBERTa-base	0.903± 0.001	0.903± 0.001
RoBERTa-large	**0.909± 0.003**	**0.909± 0.003**

Table 1. Test set results for dog whistle detection

Model	Precision	F1-score
BERT-base-uncased	0.958± 0.001	0.958± 0.001
RoBERTa-base	0.971± 0.001	0.970± 0.001
RoBERTa-large	**0.979± 0.001**	**0.979± 0.001**

Table 2. Test set results for in-group identification

While both RoBERTa models consistently outperform BERT-base across tasks, the performance gap between RoBERTa-base and RoBERTa-large is very small. Given that RoBERTa-large requires significantly more computational resources, we selected RoBERTa-base for subsequent experiments in the interest of computational efficiency and sustainability.

4.2. Generalization to Emerging Dog Whistles

A key objective of this study is to detect emerging dog whistle expressions— i.e. instances not seen during training. To assess generalization capacity, we evaluated our final RoBERTa-base classifiers on the subset of dogwhistle set aside from training. Results are reported over five runs with different random seeds in Table 3.

Task	Precision	F1-score
Dog whistle detection	**0.834± 0.004**	**0.831± 0.004**
In-group identification	**0.959 ± 0.001**	**0.950± 0.001**
Kikkisetti+2024 [7]	0.63	0.72
Xu+2022 [8]	0.81	0.78

Table 3. Performance on hidden dog whistles

As expected, both classifiers exhibit a performance drop on emerging dog whistles compared to the previous test set

12. Nevertheless, the models retain a reasonable ability to generalize to novel expressions—demonstrating the benefit of contextualized MLM pretraining.

While our model seems to outperform existing baselines in this setting, these comparisons should be interpreted with caution. The referenced works did not release their code or datasets, and their experimental setups differ in both cases. Therefore, direct comparison is not possible.

Overall, these findings confirm that our approach is effective not only in detecting known dog whistle language, but also in identifying and classifying newly emerging instances with promising accuracy.

5. Analysis

The following section examines our models' performance to uncover their strengths, limitations, and the factors influencing their successes and failures.

5.1. Confusion matrices

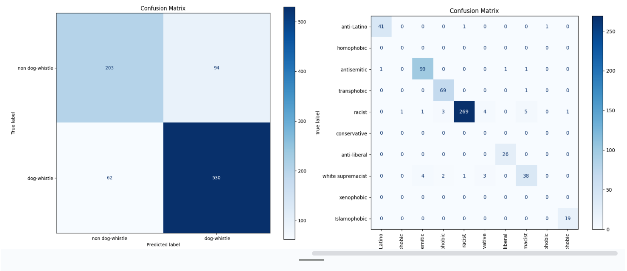
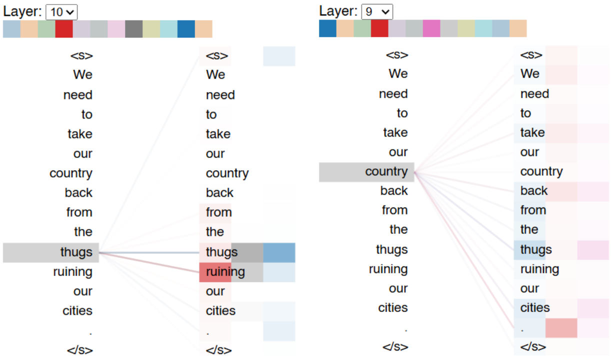


Figure 2. Confusion matrices for dog whistle detection (left) and in-group identification (right) on emerging dog whistles

Fig.2 shows confusion matrices for both classifiers on the emerging dog whistle test set.The binary classifier effectively distinguishes dog whistles but slightly tends to overpredict them. The multi-class classifier on the other hand, performs well on dominant categories like racist, antisemitic, and transphobic, but struggles with classes close in meaning such as white supremacist and racist.

5.2. Attention visualization

We used the BertViz tool [9] to visualize token-level attention patterns within our fine-tuned RoBERTa models. The visualizations show attention from a single attention head at a selected layer. Arrows indicate the direction of attention, going from the query token on the left to the key tokens on the right, with color intensity representing the attention weight. In the binary classification model, the token 'thugs' strongly attends to 'ruining' and itself, highlighting the model's focus on offensive contextual information. In the multi-class classifier instead, the token 'country' shows strong attention toward 'take back from', which the model may associate with the racist in-group.



(a) Dog whistle detection (b) In-group identification

Figure 3. Attention weights maps obtained with [9]

5.3. Failure modes

Tables 5.3 and 5.3 shows misclassified examples and possible causes.

Failure Mode	Example Phrase	Pred./True
Very specific coded language (meme, political words) not found	i have rarely encountered someone with such a firm command of ebonics bix nood	0 / 1
Language possibly offensive, but not used as such (limit in contextual information captured)	the text is black and the background is white even a colorblind person could see how black and white it is	1 / 0
Political/economic statement flagged as dog whistle	help hardworking americans weather these turbulent economic times	1 / 0
Sentences whose labels are ambiguous for a human reader (clear limit to the dataset quality)	child of an illegal immigrant who has cholera could well be the source of great damage and harm to a whole community	1 / 0

Table 4. Examples of failure modes in dog whistle classification

Failure Mode	Example Phrase	Pred./True
Confused in-groups close in meaning	its a tool for them to destroy power entities held by white western populations its a counter mechanism against judeochristian western establishments and societies destroy those and the rest will easily fall	racist / Islamophobic
Anti-liberal label for any political (also non offensive) statement (clear limit fo the dataset quality) found as racist	there are those who support hardworking american families and small businesses against those who wish to protect the status quo and big wall street banks	anti-liberal/ racist

Table 5. Examples of failure modes in in-group identification

6. Conclusion

Our work indicates that transformer-based models can be useful in detecting dog whistles and identifying their targeted groups, including for emerging expressions. Using contextual information, the models perform well in both binary and multi-class classification tasks. However, performance is is constrained by the GPT-generated nature of the dataset, and the use of fixed class definitions for in-group labeling, oversimplify the diversity of targeted communities. Despite these constraints, approach shows promise for identifying subtle and coded forms of hate speech.

Data availability

The code used in this work is available at `https://github.com/chiaraevangelisti01/DL_hate_speech`. The dataset used in this work is available at `https://huggingface.co/datasets/AstroAure/dogwhistle_dataset`.

References

[1] J. Mendelsohn, R. L. Bras, Y. Choi, and M. Sap, "From dogwhistles to bullhorns: Unveiling coded rhetoric with language models," in *Annual Meeting of the Association for Computational Linguistics*, 2023.

[2] J. Kruk, M. Marchini, R. Magu, C. Ziems, D. Muchlinski, and D. Yang, "Silent signals, loud impact: LLMs for word-sense disambiguation of coded dog whistles," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 12493–12509, Association for Computational Linguistics, Aug. 2024.

[3] D. Xu, S. Yuan, Y. Wang, A. U. Nwude, L. Zhang, A. Zajicek, and X. Wu, "Coded hate speech detection via contextual information," in *Advances in Knowledge Discovery and Data Mining* (J. Gama, T. Li, Y. Yu, E. Chen, Y. Zheng, and F. Teng, eds.), (Cham), pp. 93–105, Springer International Publishing, 2022.

[4] D. Kikkisetti, R. U. Mustafa, W. Melillo, R. Corizzo, Z. Boukouvalas, J. Gill, and N. Japkowicz, "Using LLMs to discover emerging coded antisemitic hate-speech in extremist social media," *arXiv e-prints*, p. arXiv:2401.10841, Jan. 2024.

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

[6] 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," 2019.

[7] D. Kikkisetti, R. U. Mustafa, W. Melillo, R. Corizzo, Z. Boukouvalas, J. Gill, and N. Japkowicz, "Using llms to discover emerging coded antisemitic hate-speech in extremist social media," 2024.

[8] D. Xu, S. Yuan, Y. Wang, A. U. Nwude, L. Zhang, A. Zajicek, and X. Wu, "Coded hate speech detection via contextual information," in *Advances in Knowledge Discovery and Data Mining* (J. Gama, T. Li, Y. Yu, E. Chen, Y. Zheng, and F. Teng, eds.), (Cham), pp. 93–105, Springer International Publishing, 2022.

[9] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Florence, Italy), pp. 37–42, Association for Computational Linguistics, July 2019.