Hertie School

Data Science Lab

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

Lino Zurmühl

Supervisor: Prof. Simon Munzert, PhD

A report submitted in fulfilment of the requirements of
the Hertie School for the degree of
Master of Science in *Data Science for Public Policy*

Word count: 7451
GitHub Repository: Master Thesis Dog Whistle

April 29, 2025

Abstract

This thesis explores the automated detection of "dog whistles", coded language used to signal often harmful messages to specific groups while appearing innocuous to others. This phenomenon, increasingly prevalent online as "algospeak" to evade moderation, presents a significant challenge due to its context-dependent nature. This project utilized the Silent Signals dataset, employing confident learning and manual review to mitigate labeling noise. I fine-tuned RoBERTa and Gemma 2, evaluating their performance on detecting right-wing dog whistles in Reddit posts. While RoBERTa was fully fine-tuned, the Gemma 2 model was adapted through Low Rank Adaptation. After the fine-tuning process, RoBERTa showed stronger in-distribution performance, while Gemma 2 demonstrated better generalization on manually labeled external data. Both models underperformed human annotators and proprietary systems, highlighting the task's complexity and the limitations of current methods and data quality. The findings underscore the need for refined datasets and context-aware models to effectively address subtle, coded communication in digital spaces.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

On January 20, 2025, during President Donald Trump's second inauguration celebrations, Elon Musk concluded his speech with a gesture widely interpreted as a Roman or Nazi salute. The act immediately sparked widespread controversy. Critics condemned it as a covert endorsement of far-right ideologies, while supporters dismissed the response as an overreaction to a misunderstood motion. Regardless of intent, the incident reignited global debates about the power of symbolic communication in public discourse and the difficulty of distinguishing between overt statements and covert signaling. It illustrated the very nature of dog whistles: ambiguous gestures or phrases that carry seemingly innocuous meanings to a general audience but resonate with a deeper, often ideological message for an intended subgroup. This thesis builds on precisely that challenge, detecting dog whistles in language, and explores how modern natural language processing (NLP) can be used to uncover such covert messages, particularly in online spaces where traditional forms of moderation fall short.

Dog whistles, coded language with layered meanings, have long been employed in political rhetoric to signal controversial or exclusionary views without explicit articulation. In recent years, their use has spread beyond elite political circles and into everyday online discourse. Social media platforms, forums, and comment sections now host a wide range of coded language used by political influencers, fringe communities, and anonymous users alike. While some of these messages are crafted to maintain plausible deniability among a broader audience, others are designed to evade detection by automated content moderation systems. This form of strategic ambiguity, referred to as "algospeak", enables the spread of hate speech, conspiracy theories, and extremist ideologies while avoiding algorithmic flagging, demonetization, or platform bans (Magu et al., 2017; Witten, 2023).

The task of dog whistle detection is therefore not only a linguistic challenge but also a crucial component of modern content moderation strategies. Unlike traditional hate speech detection, which often relies on clear-cut semantic cues, dog whistle detection requires identifying terms or phrases whose meaning depends heavily on context, community norms, and cultural knowledge (Witten, 2023; Khoo, 2017; Henderson and McCready, 2018; Haney-López, 2014; Sayeed et al., 2024). These messages may appear harmless in isolation but take on coded significance within certain discourse communities. For content moderation systems to effectively address such cases, they must move beyond keyword filtering and adopt context-aware, adaptive models that can understand subtle variations in usage. Automated detection of dog whistles could help platform moderators identify content that would otherwise evade scrutiny, offering a proactive tool in the fight against covertly harmful speech online.

Recent advancements in large language models (LLMs) provide new tools for approaching this task. Models like RoBERTa and Gemma 2, when fine-tuned on domain-specific data, can learn contextual embeddings that enable them to distinguish between literal and coded meanings.

These transformer-based architectures have shown strong performance on a range of classification tasks, including hate speech and offensive language detection(Caselli et al., 2021; Mozafari et al., 2020; Saleh et al., 2023). However, their ability to detect dog whistles, where meaning is intentionally obscured, has yet to be fully explored. More importantly, the success of such models depends not only on their underlying architecture but also on the quality of training data, the presence of label noise, and their ability to generalize across domains.

This thesis investigates the task of dog whistle detection using fine-tuned language models, with a focus on their application in content moderation. Particularly it tries to answer the question: How effective are fine-tuned language models in accurately classifying implicit hate speech characterized as right-wing dog whistles Reddit posts? It evaluates the performance of RoBERTa and Gemma 2 across both in-distribution and out-of-distribution data, drawing on a dataset of political informal online comments (Kruk et al., 2024). It further examines the impact of noisy labeling, architectural differences, and dataset composition on model outcomes. In doing so, the thesis contributes to the broader goal of building more context-sensitive moderation systems capable of addressing the increasingly sophisticated ways in which language is used to evade scrutiny and cultivate harmful ideologies in digital spaces.

Chapter 2

Literature Review and the Definition of a Dog Whistle

2.1 Dog Whistle Detection in the Literature

The challenge of detecting coded communication, particularly political "dog whistles", has gained increasing attention within Natural Language Processing (NLP)[1]. Dog whistles represent a subtle form of language designed to convey a secondary, often harmful, meaning to a specific in-group while appearing innocuous to a broader audience, thereby evading simple keyword-based detection and content moderation. Early work by Magu et al. (2017) focused on "hate codes" used to circumvent platform policies, highlighting the need for methods beyond explicit hate speech detection.

Recent research has explored the complexities of implicit hate speech, which shares characteristics with dog whistles in its subtlety and reliance on context. Hartmann et al. (2025) note that research on commercial content moderation APIs is limited, particularly concerning implicit hate speech. Datasets like ToxiGen Hartvigsen et al. (2022) were created using large language models (LLMs) to generate examples of implicit toxicity for model training and evaluation. Although both implicit hate speech and dog whistles involve non-explicit harmful language, dog whistles specifically rely on coded expressions designed to convey a secondary meaning to a narrow in-group while appearing innocuous to a broader audience, offering plausible deniability (Mendelsohn et al., 2023). While work exists on representing dog whistles using models like BERT Hertzberg et al. (2022), the specific task of disambiguating whether a term is used in its coded sense or its neutral sense remains a significant hurdle.

Recognizing this, Kruk et al. (2024) introduced the "Silent Signals" dataset, leveraging LLMs for the word-sense disambiguation of dog whistles identified in Mendelsohn et al. (2023) glossary. This dataset was explicitly created with the goal of facilitating the training and fine-tuning of models for tasks like hate speech detection and identifying emergent dog whistles.

While the advantage of fine-tuning LLMs for related tasks like abusive language detection has been demonstrated (Nguyen et al., 2023; Sen et al., 2024), including work on smaller or "tiny" LLMs, a gap exists in applying these techniques specifically to the Silent Signals dataset for dog whistle detection. Kruk et al. (2024) focused on using LLMs for dataset creation via disambiguation, and while they evaluated LLMs' zero-shot detection capabilities, they did not report on fine-tuning models using the resultant dataset. Other studies have fine-tuned models like Llama 2 or tiny LLMs for general abusive text or hate speech, but not specifically for the

[1]Note on the machine learning literature: Due to the norms of academic publishing in machine learning, many works appear as conference papers, technical reports, or preprints, and may not undergo rigorous peer review in the traditional journal sense.

nuanced task of dog whistle detection using the specialized Silent Signals resource.
This thesis addresses this gap by fine-tuning state-of-the-art language models, specifically Gemma 2 and RoBERTa, on the Silent Signals dataset. By directly utilizing the dataset curated by Kruk et al. (2024) for this purpose, I aim to evaluate the effectiveness of fine-tuning for improving the detection of coded dog whistles, contributing to the ongoing effort to mitigate this form of harmful online communication.

2.2 Defining Dog Whistles in Online Contexts

The exact definition of dog whistles has been an ongoing debate in the political science and linguistics literature. The term itself first appears in a political context in 1988, according to Safire (2008), who cites Richard Morin of the Washington Post using it to describe subtle changes in poll questions that result in large shifts in respondents' answers (Morin, 1988). The term was then adopted to describe the phenomenon of political coded language used by politicians to appeal to specific audiences without explicitly saying so (Witten, 2023; Khoo, 2017; Henderson and McCready, 2018; Haney-López, 2014; Sayeed et al., 2024). Most authors use specific dog whistle examples uttered by George W. Bush, Donald Trump, Richard Nixon, or Ronald Reagan as their springboard for defining the term. Here, the focus on US politicians is apparent and shows that most of the research yet has been done on defining dog whistles in the Anglo-American context. While the exact definition varies between approaches, the main overarching characteristics are the duality of meaning, where there is a general overt message and a covert hidden message, and the concealment of this hidden message.

How do Dog Whistles Function?

Henderson and McCready (2018) highlights four main theoretical perspectives regarding the mechanisms underpinning dog whistles: Conventional Implicature (CI), Inferentialist, Game-Theoretic, and Mixed Views. The Conventional Implicature view suggests that dog whistles convey hidden meanings conventionally encoded within language, making them semantically or pragmatically ambiguous (Stanley, 2015). In contrast, the Inferentialist perspective, represented by Khoo (2017), argues that dog whistles function through listeners' inferential processes, activated by their pre-existing beliefs and stereotypes rather than by any intrinsic semantic ambiguity in the words themselves. This perspective emphasizes the listener's cognitive contribution rather than the speaker's explicit coding.
The Game-Theoretic perspective focuses on strategic interaction, utility, and incentives, emphasizing that dog whistles are strategically used by speakers to maintain plausible deniability while communicating sensitive content to particular subgroups (Sayeed et al., 2024). Finally, the Mixed View integrates aspects from Inferentialist and Game-Theoretic perspectives, proposing that dog whistles operate via a combination of inference based on audience beliefs and strategic linguistic design related to the speaker's persona and context (Henderson and McCready, 2018). Witten (2023) offers a detailed typological model, emphasizing elements such as intentionality, concealment, and distinct audience design. Although Witten's typology primarily describes dog whistles structurally rather than mechanistically, it aligns closely with the Mixed View in considering both speaker intent and audience inference.

Political And Social Effects

While the above-mentioned authors go into the detail of how dog whistles work in a linguistic philosophical sense, the effect of those dog whistles on political and social spheres and the effect of political and social changes on the definition of the term is only briefly mentioned.

Those effects are crucial in understanding the nature of such coded language.
Mendelberg (2001) argues that dog whistles emerged after the Jim Crow era, when the societal acceptance of explicit racism diminished through the civil right movement. Using coded language gave politicians plausible deniability while appealing to racial sentiments. Using dog whistles like "law and oder" as to voice latent grievances and signal opposition to the civil rights movement. The same rhetoric can be seen today in Trumps use of "law and order" as an oppositional statement against the Black Lives Matter movement (Drakulich et al., 2020). Research suggests that such coded appeals can significantly impact voter behavior and polarization. Dog whistles work by activating latent biases or resentments within targeted subgroups, potentially swaying opinions or mobilizing support for certain policies or candidates more effectively than overt statements might, precisely because they operate implicitly (Wetts and Willer, 2019; Sayeed et al., 2024). This activation of prejudice, particularly racial prejudice, can contribute to affective polarization—increasing hostility towards opposing groups, even if it doesn't uniformly shift ideological stances (Iyengar et al., 2012; Suhay et al., 2018). Historically, much of the research has focused on politicians as the addressers and the broader public, or specific voter segments, as the addressees (Haney-López, 2014; Albertson, 2015). This focus stems from the origins of the term in political strategy, where elites required coded ways to communicate with parts of their constituency without alienating others or explicitly violating evolving social norms, such as those concerning race (Mendelberg, 2001). The primary goal was political influence and electoral advantage within the constraints of public discourse. While dog whistles effect the political and social behavior, conversely the political context also shapes the role of dog whistles. In the past research has focused on dog whistles being used by politicians to avoid public scrutiny. But with the rise of social media, the internet has become one of the most important places for political discourse. Through internet virality, non-traditional actors like political influencers can now shape public opinion and are privy to public scrutiny online. Here the plausible deniability towards the other community members is also present, but a second goal of dog whistles emerges. With open racism and hate speech mostly banned or at least regulated on popular online platforms the use of dog whistles is a way to bypass automated or moderator based restrictions (Magu et al., 2017). This represents a new form or primary function of dog whistles: online "algospeak" or coded language specifically designed for algorithmic evasion.Users substitute potentially flagged terms related to hate speech, controversial topics, or targeted communities with seemingly benign words, misspellings, or symbols to avoid content moderation systems, demonetization, or account suspension, while still conveying the intended meaning to their in-group (Witten, 2023).
Both Mendelsohn et al. (2023) and Kruk et al. (2024) extend the concept of dog whistles to encompass both political and online communication contexts. Mendelsohn et al. (2023) differentiate terms by register, contrasting 'formal/offline' language, such as political speeches, with 'informal/online' language derived from internet culture. In parallel, Kruk et al. (2024) apply a general definition of dog whistles as coded language with a secondary meaning and construct a dataset that draws from both U.S. Congressional records (formal) and Reddit (informal). Their work explicitly acknowledges the role of dog whistles in evading hate speech detection, reflecting the evolving nature of this linguistic phenomenon in digital environments. However, while they incorporate emerging forms of "algospeak" into their dataset, they do not distinguish this type of algorithm-evasive language as a distinct category in the informal dataset. Instead, traditional political dog whistles and contemporary internet-coded speech are treated as a unified phenomenon, potentially overlooking important functional and contextual differences between the two.

Chapter 3

Methods

3.1 Initial Data Description

Silent Signals Data Set

Kruk et al. (2024) published the "Silent Signals"[1] dataset in conjunction with their ACL paper. Here, the researchers used the best performing method of an ensemble approach for "Dog Whistle Disambiguation" with GPT-4 to curate a dataset of 16,550 dog whistles. This dataset was published with the goal of "be[ing] used to training and/or finetuning [...] for tasks ranging from hate speech detection to emergent dog whistle identification" (Kruk et al. (2024)), aligning with the goal of this thesis. After their experiments with different prompt setups, the researchers achieved a precision for the positive class of 96.2%. This was achieved by adding the Wikipedia definition of the specific dog whistle and the successive prompting of the same example. Using this method, Kruk et al. labeled a random sample from the Potential Instance dataset curated based on the keyword glossary provided by Mendelsohn et al. (2023). The instances that were positively labeled by GPT-4 were collected, and 400 of them were manually reviewed by the researchers. Here the researchers found a precision of 85.3% after human review.
Overall, the dataset is divided into two different categories of dog whistles: formal and informal. Following the definition provided by Mendelsohn et al., Kruk et al. distinguishes between two registers of dog whistles. Formal or offline dog whistles are those that originate in offline contexts and are more likely to be employed by mainstream political elites. Informal or online dog whistles, on the other hand, emerge from Internet culture and are generally not used in conventional political discourse. In the Silent Signals dataset, formal dog whistles account for 20.2% and informal dog whistles account for 79.8%. For the purposes of this thesis, I will focus on informal dog whistles due to the larger distribution of positive instances for training, as well as for potential online use cases of a dog whistle detection model discussed in Chapter 2.

Null Instances

Training a binary classifier requires both negative and positive examples. The published Silent Signals data consists only of posts that have been labeled as containing a dog whistle with high accuracy by GPT-4. Negative examples are missing from this published dataset. To address this, this paper includes the Reddit posts that were randomly sampled by Kruk et al. from the Potential Instances dataset, but were not considered containing a dog whistle by GPT-4. Since

[1] https://huggingface.co/datasets/SALT-NLP/silent_signals

the original method is optimized for positive class accuracy, the negative texts are expected to be more noisy and contain more false negatives than the positive examples in the Silent Signals dataset. In order to mitigate the labeling errors, a combination of statistical methods and manual label checking is employed and is further explained in the Section 3.2. Before this preprocessing step, the data contains 34,212 text instances containing a keyword from 299 different dog whistle roots. These posts come from 39 Subreddits and contain 18 ingroups.[2]

Silent Signals Detection Data Set

In addition to the Silent Signals dataset, Kruk et al. also published 101 manually selected texts to test the zero shot detection capabilities of LLMs[3]. For these texts, different LLMs were prompted if a dog whistle was present, paired with a definition of the dog whistle. Again, the best performing model was GPT-4. All positive examples (51) are Reddit posts, but the negative examples (50) also contain congressional texts. This data set will be used in Chapter 4 to compare the models fine-tuned on the combined preprocessed dataset with the off the shelf LLMs used by Kruk et al..

3.2 Data Cleaning and Exploratory Data Analysis

Data cleaning with Confident Learning

To address the potential presence of mislabeled instances in the Silent Signals dataset, I used Confident Learning (CL), a framework introduced by Northcutt et al. (2021). This approach is designed to identify and correct labeling errors in datasets by exploiting the confidence of a trained classifier along with the inherent structure of the data.
In the first step, BERT embeddings (Devlin et al., 2019) were computed for the informal texts from the combined Null and Silent Signals dataset. These embeddings were then used to train a logistic regression classifier[4] with 5-fold cross-validation to ensure out of sample predictions. Next, the true label distribution of the data set was estimated. By analyzing the joint distribution of observed labels and predicted confidence scores, we can infer the probability that the true label of a given instance is different from its assigned label. Specifically, we estimate the probability $P(Y \neq \hat{Y} \mid X)$, where Y is the true label, $\hat{Y}$ is the observed label, and X represents the features of the instance. This estimation relies on the framework introduced by Angluin and Laird (1988), where the noise rate η is modeled as $P(Y \neq \hat{Y}) < \frac{1}{2}$, ensuring that the data is statistically distinguishable from noise. By estimating the classifier's confidence scores, we can approximate the probability that a given instance belongs to a misclassified subset of the data, denoted as

$$L_{\text{mislabel}} = \{X_i : P(Y \neq \hat{Y}|X_i) > \varepsilon\}$$

for the threshold of $\varepsilon > 0$. Labeling errors are then identified by examining cases where the classifier predicts a class with high confidence that contradicts the observed label. For example, if the model predicts with high confidence that an instance labeled 0 should be 1, this discrepancy suggests a possible mislabeling. The assumption underlying this process is that such high-confidence disagreements are indicative of noisy or incorrect labels.
After running the CL pipeline on the raw data, 6,525 potential mislabels were identified. Of these, 52.67% had the original label 1 and 47.33% had the label 0. With a high class imbalance

[2]Note that these posts contain a keyword associated with a dog whistle, but GPT-4 recognized it as being used in an uncoded context.

[3]https://huggingface.co/datasets/SALT-NLP/silent_signals_detection

[4]The model choice here is due to the high efficiency and low resource requirements of a logistic regression model, especially with K-fold cross-validation.

in favor of the null class, as well as the focus on precision for the positive class in Kruk et al.'s disambiguation method, we would expect a larger number of negative items to be flagged as potentially mislabeled. The Chapter 5 discusses this problem in more detail.

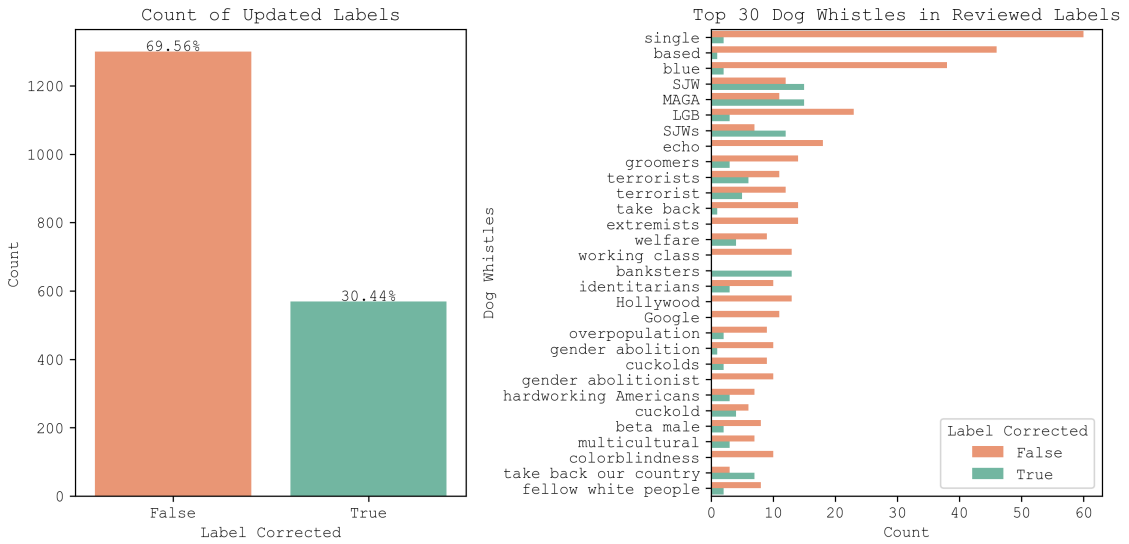

Figure 3.1: Distribution of true negative labels and corrected negative to positive labels (N = 2000). The left hand figure looks at the overall distribution and the right hand figure looks at the top 30 dog whistle in the potential mislabels and their manual check outcome.

To further reduce class imbalance, 2,000 negative instances flagged as potentially positive were manually reviewed. 131 texts were removed because they were in a language other than English, were incoherent, or could not be definitively assigned to either class. Figure 3.1 shows the proportion of corrected labels and the top 30 dog whistles present in the reviewed by changed label. Of the remaining 1,869 items, 69.56% were correctly labeled by GPT-4 as not containing a dog whistle. This is expectedly lower than the positive class precision of 85.4% reported by Kruk et al. (2024) for their N = 400 positive sample[5]. Looking at the right-hand figure, texts containing potential dog whistles based on everyday words like "single" or "based" are mostly correctly labeled, and instances containing "banksters," "SJW," or "MAGA" were mostly incorrectly labeled as not containing a dog whistle.

After manual review, the 1,869 reviewed instances were returned to the corpus. The remaining unseen potential mislabels were removed from the training data. Their labels could not be confidently verified and retaining them risked introducing more noise into the model. In addition to the mislabels, this preprocessing step included removing 165 posts that were considered to be near duplicates (highly repetitive moderator messages or live tickers).

To test the effectiveness of the preprocessing step, the logistic model based on the BERT embeddings was rerun on the processed data, with 10% held back as a validation set, and compared to the model trained on the original data.

Figure 3.2 and Table 3.1 show a substantial increase in performance for all metrics after data cleaning. In particular, the F1 score of the positive class increases by 15 percentage points. Eliminating the more noisy instances as well as re-incorporating the manually checked instances proved to be effective.

[5]We are looking at already flagged potential mislabeling instances, not a random sample as in Kruk et al.

Table 3.1: Performance comparison of logistic regression model before and after label updates.

Model	Class	Precision	Recall	F1-Score	Support
	0 (Non-Dog Whistle)	0.81	0.91	0.86	3448
	1 (Dog Whistle)	0.63	0.42	0.50	1264
Original	Accuracy			0.78	4712
	0 (Non-Dog Whistle)	0.88	0.93	0.90	3231
	1 (Dog Whistle)	0.72	0.59	0.65	1002
Updated	Accuracy			0.85	4223

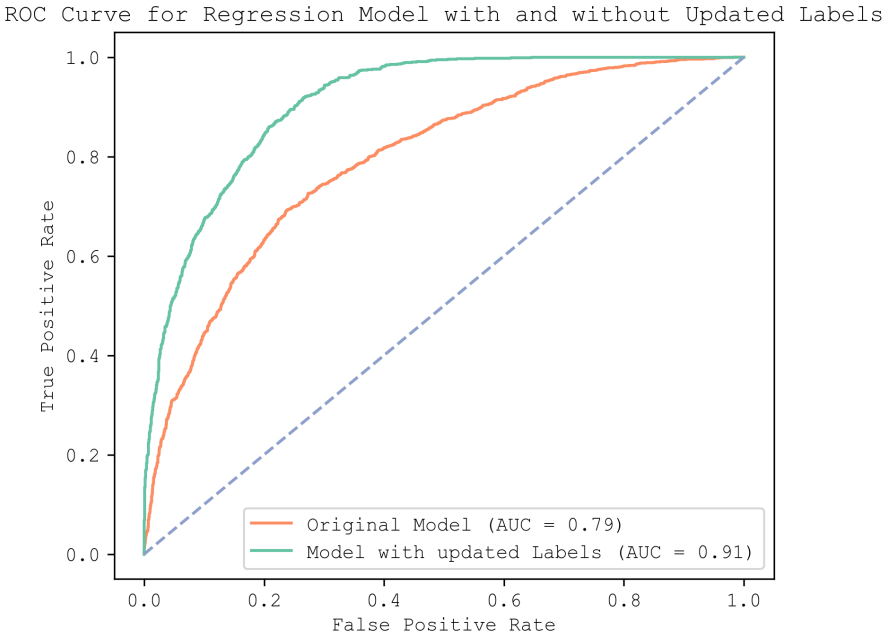

Figure 3.2: Area under the curve for the positive class before and after data cleaning of the logistic regression model.

Exploratory Analysis of the Full Dataset

Looking at the fully processed dataset, we got 42,323 texts with 10,021 positive dog whistle labels and 32,302 negative dog whistle labels. The posts come from 45 subreddits, spanning from June 3, 2008 to December 2, 2023. In Figure 3.3 we can see the distribution of posts over time. Two spikes can be observed: The first after 2016 and the second after 2021. The spike for both classes appear to be proportional in 2016, while the 2021 spike is disproportionately driven by the null class[6].

The frequency of dog whistles keywords from Mendelsohn et al. (2023) for the two classes is also not evenly distributed. When looking at Figure 3.4, for the positive label, "social justice warrior(SJW)", a term used to describe left leaning political activists by the right, is the most frequent dog whistle root preset in the data set with nearly 600 posts. "Cuck"[7] and "Centipede"[8] follow with around 280 posts for the positive class. Racial and antisemitic dog

[6] Also see 3.4 in Appendix A.
[7] Used as a derogatory code for liberal.
[8] Code for Trump supporter.

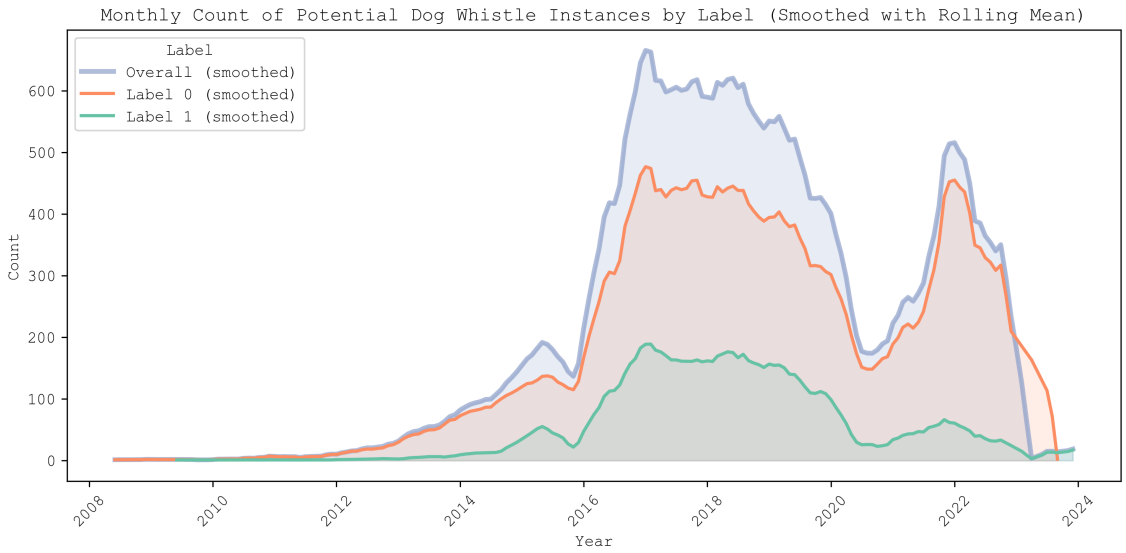

Figure 3.3: Number of Reddit posts over time by label with moving average of a six month span, showing two spikes for 2016 and 2021.

whistle are also present in the top 15 roots with terms like "globalist"[9], "illegal immigrants" and "illegal aliens". In contrast the right panel shows for the most common key word root "lifelong bachelor". This is due to the dog whistle "single"[10] inherent to this root being frequently used in colloquial language without political context. "Google"[11] and "based"[12] can also be frequently found in general discourse, explaining their prevalence in the null class.

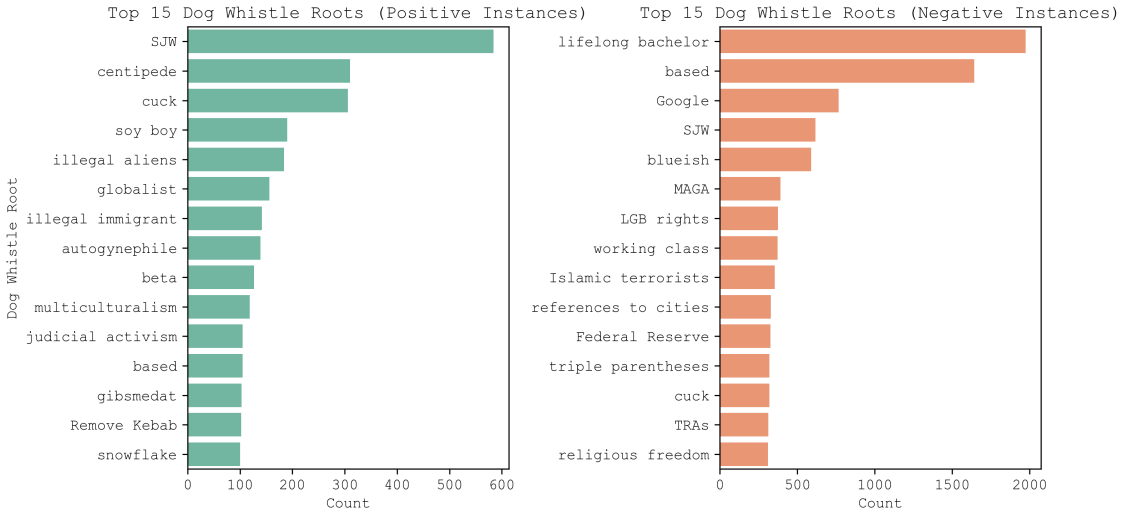

Figure 3.4: Top 15 dog whistle roots per label, showing most prevalent keywords for each class.

Nevertheless, there are top 15 overlapping keyword in both classes. A term like "SJW" appears over 500 times in the negative class, and "cuck" appears roughly 300 times. This overlap

[9] Code for Jew.
[10] Homophobic code calling someone single/not married to suggest that they are homosexual.
[11] Code for a black person.
[12] Someone or something that is considered to be accepted as alt-right.

underlines the key challenge of dog whistle detection: many of these roots occur both in hateful and in neutral contexts. The same lexical token can be harmless or weaponized depending on its surrounding words, highlighting the dependency on context aware methods for detecting dog whistles.
For training the data set was split at random into 38,090 instances (90%) for training and 4233 (10%) instances for validation with a class distribution of 76% (null) to 24% (positive).

3.3 Dog Whistle Classification with Language Models

Detecting dog whistles in text is, as mentioned in section 2.2, a highly context-dependent task. Classical natural language processing methods such as bag-of-words coupled with a linear model would not be able to adequately model the surrounding words that influence the meaning of the keyword associated with a dog whistle. Kruk et al. showed that using commercial off-the-shelf LLMs with zero shot classification can show promising results. Models like GPT-4 can archive high precision in such a context dependent task trough their transfer learning abilities. Although LLMs show promising results, they have two major drawbacks: First, LLMs show worse performance on zero shot classification tasks that are out of their training distribution (Li and Flanigan, 2023). This is critical for a task as volatile as dog whistle classification, where new dog whistles appear and old ones disappear quickly. Second, using a model like GPT-4 can be recourse-intensive from an environmental perspective (Kaack et al., 2022) and makes a detection pipeline dependent on closed-source models with very limited control over model biases or outputs. In the next two subsections, I will propose two alternative approaches made possible by the combined dataset described in the section above. With this dataset, context-aware language models can be fine-tuned for the specific task of dog whistle detection.

RoBERTa finetuning

Fine-tuning a BERT-based model has been shown to be effective in various classification tasks and has been particularly popular for hate speech classification (Caselli et al., 2021; Mozafari et al., 2020; Saleh et al., 2023). BERT-based models have also been used to detect implicit hate speech (Ocampo et al., 2023) and the semantic change of dog whistles over time (Boholm and Sayeed, 2023). Building on this foundation, this thesis uses RoBERTa by Liu et al. (2019), a robustly optimized variant of BERT, and fine-tunes the model to the dog whistle dataset. The transformer-encoder-only architecture allows BERT-based models to use their pre-trained "understanding" of language to classify text based on context-rich embeddings. Central to this architecture is the attention mechanism introduced by Vaswani et al. (2017), which allows the model to weigh the importance of different words in a sequence relative to each other, regardless of their distance, allowing for nuanced contextual understanding. Another advantage of the encoder-only architecture is the bidirectional context embeddings produced through masked language modeling, the model not only to "see" context from left to right, but can look both forward and backward in the word string (Devlin et al., 2019).
The fine-tuning process included hyperparameter tuning by grid search for 18 paths. Here, two variants of RoBERTa, the large (355M parameters) and the base (125M parameters) model, were tested along with tuning the learning rate and weight decay on 50% of the training data. In order to train a model that is reliably detecting dog whistles the F1 score of the positive class was chosen as the main classification metric for model performance. To address the high class imbalance in the dataset class weights were added to the cross entropy loss.
The final model was then trained for three (early stopping) epochs with a learning rate $=$

2e-05, no weight decay, and the large variant on four A100 GPUs. After training, a final evaluation was run on the held-out validation set (10% of the total data).

Gemma 2 Finetuning with Low Rank Adaptation

With the rise of generative LLMs, their adaptability to specific tasks also increased. Relying on the broad linguistic understating of pre-trained generative models, new models could be fine-tuned to specific subdomains with limited data. While zero-shot classification can be an easy way to exploit the transfer learning capabilities of generative LLMs, fine-tuning on top of the basic understanding can yield robust results (Davidson and Chae, 2025). Adapting a large generative language model to your training data is a time- and resource-intensive process that has made it accessible only to large corporations and research clusters. However, with increasing efficiency of distillation, the performance of Small Language Models (SLM)[13] can show similar performance to their larger counterparts (Gu et al., 2024) and have been used by researchers as an efficient alternative to LLMs (Mo et al., 2024; Nguyen et al., 2023; Sen et al., 2024).
Building on recent developments in efficient generative models, this thesis fine-tunes the 2 billion parameter Gemma 2 model by Team Gemma et al. (2024), a decoder-only Transformer optimized for performance at small scale. The model introduces several architectural innovations, including alternating attention layers, grouped query attention for memory efficiency, rotary position embeddings for improved handling of token distance, and gated GeLU activations for richer representation learning. The 2b model is pre-trained via full-distribution knowledge distillation from a larger teacher model.

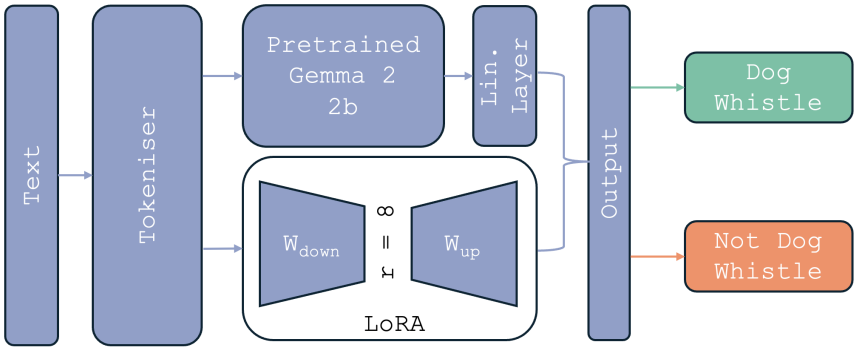

Figure 3.5: Finetuning process of Googles Gemma 2 model with 2 billion parameters through LoRA with Rank = 8. The parameters in linear layer and pretrained Gemma 2 model are frozen and weight matricides injected through LoRA into specific modules.

In contrast to the smaller RoBERTa model, fully finetuning the Gemma model was out of scope for this thesis. Utilizing Low Rank Adaptation for finetuning reduces the amount of parameters being trained and can produce similar performance as full finetuning (Hu et al., 2021). By freezing the pretrained parameters of the model and adding ranked decomposition matricides into specific layers of the model, LoRA uses only those matricides for training.
This is based on the assumption that the required update to a weight matrix, ΔW, often lies in a much lower-dimensional subspace. Rather than modifying the full weight tensor W

[13] Here the literature has no clear definition of what exactly is the parameter number that distinguishes between a "large" and a "small" language model (Naveed et al., 2024). For the sake of this thesis I define an SLM as having fewer than 7 billion parameters.

directly, LoRA freezes W and injects trainable parameters via two smaller matrices:

$$W' = W + W_{\text{up}} \times W_{\text{down}}$$

Here, $W \in \mathbb{R}^{d \times d}$ remains unchanged during fine-tuning. The matrix $W_{\text{down}} \in \mathbb{R}^{r \times d}$ projects the d-dimensional activations down to a rank-r subspace, and $W_{\text{up}} \in \mathbb{R}^{d \times r}$ lifts them back to the original dimension. Their product $W_{\text{up}} \times W_{\text{down}}$ thus serves as a low-rank approximation of ΔW.
By expressing the update as a product of W_{down} and W_{up}, the number of trainable parameters falls from d^2 in a full-rank update to just 2dr. Sen et al. demonstrated that for hate speech classification this utilization of LoRA outperformed other adaption methods for SLMs.
For the dog whistle detection model a fully connected linear neural network was added as a classification head. This addition adapted the model from its original task (next token prediction) to being a classification model that outputs a probability distribution over binary labels, indicating the presence or absence of a dog whistle in the input text. The whole finetuning setup can be seen in Figure 3.5 and targeted layers with LoRA in Appendix A Table A.1.
Next to the LoRA adaptation, hyperparameter tuning by grid search for 18 paths was run. Here the rank of the LoRA matricides r, learning rate and weight decay were tuned on 30% of the training data. Again F1 score was chosen as the classification metric for best model performance and class weights were added to the cross entropy loss function. To address the generative nature of the pretrained Gemma model, the training and test texts were wrapped in a prompt containing an instruction and the definition of a dog whistle similar to the process shown in Kruk et al.[14]. The final model was then trained for three epochs (early stopping) with a learning rate = 0.0001, weight decay = 0.01, and r = 8 on four A100 GPUs. After training, a final evaluation was run on the held-out validation set (10% of the total data).

[14]The prompt structure can be seen in Appendix A.

Chapter 4

Results

After training, both models were evaluated on two datasets: the internal holdout set—drawn as a random sample from the same distribution as the training data and an external test set consisting of 101 annotated examples for dog whistle detection provided by Kruk et al..
In the first section, both models are compared in terms of their performance on the in-distribution holdout set to assess how well they generalize to unseen but similar data. The second section evaluates their performance on the external detection data set, offering a comparison with commercial out-of-the-box LLMs and overall generalization performance of the fine-tuned models.

4.1 In-Distribution Performance

After training on the 38,090 training instances, both models were evaluated on the 4,233 holdout instances. Both models show promising results in the distribution. Due to the class imbalance of about four to one in favor of the null class, a naive model that only classifies the texts as not containing a dog whistle would get an accuracy of about 0.76. Compared to this baseline, both models show substantial improvements. The fine-tuned Gemma 2 model classifies 88% of the text correctly, and the fine-tuned RoBERTa model surpasses this with 90% of the items correctly labeled. To accurately measure performance in the minority class (texts containing a dog whistle), looking at the F1 score is more reliable. As the harmonic mean of precision and recall ensures that both are true positives, the model has balanced performance across classes. For the purpose of a model for the detection of dog whistles, the focus lies on the positive class. Again, the fine-tuned RoBERTa model outperforms the

Table 4.1: Classification Report for RoBERTa on In-Distribution Validation Set

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0 (Non-Dog Whistle)	0.98	0.89	0.93	3231
1 (Dog Whistle)	0.73	0.93	**0.82**	1002
Accuracy		**0.90**		4233

fine-tuned Gemma 2. While Gemma 2 shows a more balanced trade-off between precision and recall, it is overall less effective at detecting dog whistles in the validation set with an F1 score on 0.76. RoBERTa reports a F1 score of 0.82. It identifies 93% of all dog whistles in the validation set, and 73% of its positive predictions are correct, indicating a tendency to over-predict the positive class and produce more false positives.

Table 4.2: Classification Report for Gemma 2 on In-Distribution Validation Set

Class	Precision	Recall	F1-Score	Support
0 (Non-Dog Whistle)	0.93	0.91	0.92	3231
1 (Dog Whistle)	0.73	0.78	**0.76**	1002
Accuracy		**0.88**		4233

To further assess the classification performance of both models, ROC curves were generated in Figure 4.1 based on their predictions on the holdout validation set. These curves plot the true positive rate (TPR) against the false positive rate (FPR) at varying thresholds and provide a threshold-independent view of model performance. The Area Under the Curve (AUC) offers a single metric summarizing how well a model distinguishes between the two classes. The fine-tuned RoBERTa model achieves an AUC of 0.97, indicating strong separability between texts that do and do not contain dog whistles. Its curve remains consistently above that of the Gemma 2 model, particularly in regions where the false positive rate is low. The Gemma 2 model reaches an AUC of 0.93, which still reflects robust performance but falls short of RoBERTa's across most thresholds. This performance gap aligns with the previously observed differences in recall and F1 score.

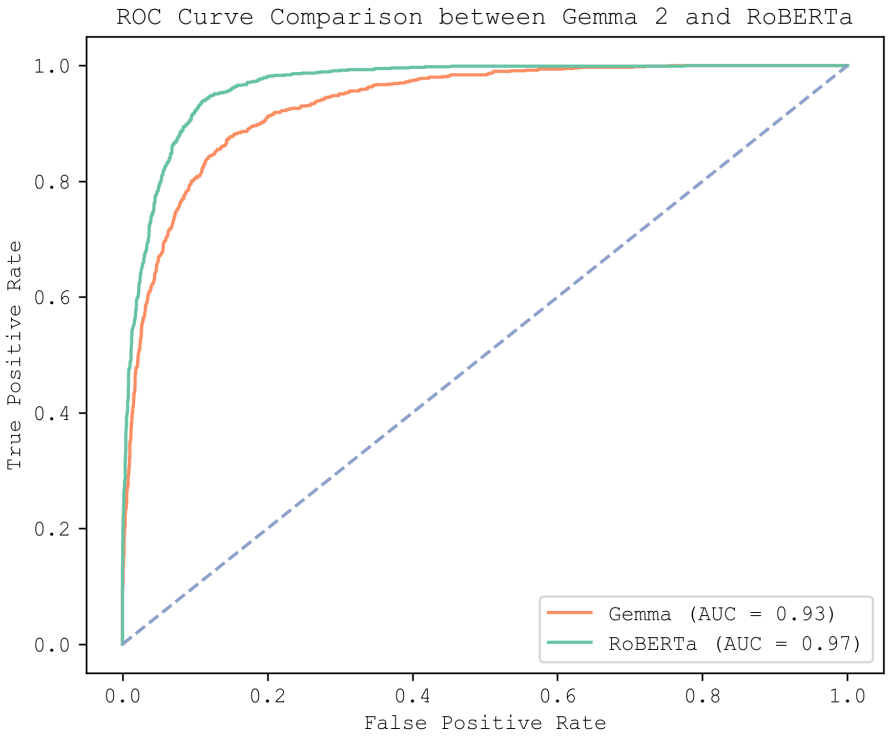

Figure 4.1: Area under the curve for the positive class of Gemma 2 and RoBERTa fine-tuned Models.

4.2 Comparison to GPT-4 with Out-of-Distribution Data

By publishing their dog whistle detection dataset, Kruk et al. provided an opportunity to test the fine-tuned models on data that is evenly distributed across the two classes, as well as compare their performance to out of the box models like GPT-4 with zero shot classification. After running both fine-tuned models on the 101 instances, we can see a substantial decrease in performance in both models. Fine-tuned RoBERTa archives an accuracy of 46% and lies under the naive baseline of only predicting null for all texts. While fine-tuned Gemma 2 beats the naive baseline, it only correctly classifies 64% of the texts. Going into more detail in the confusion table in Figure 4.2, RoBERTa only correctly detects one of the 51 dog whistles present in the texts. In contrast, the Gemma 2 model demonstrates more robust generalization. It correctly identifies 19 out of 51 dog whistle cases, but still falls short of the in-sample performance.

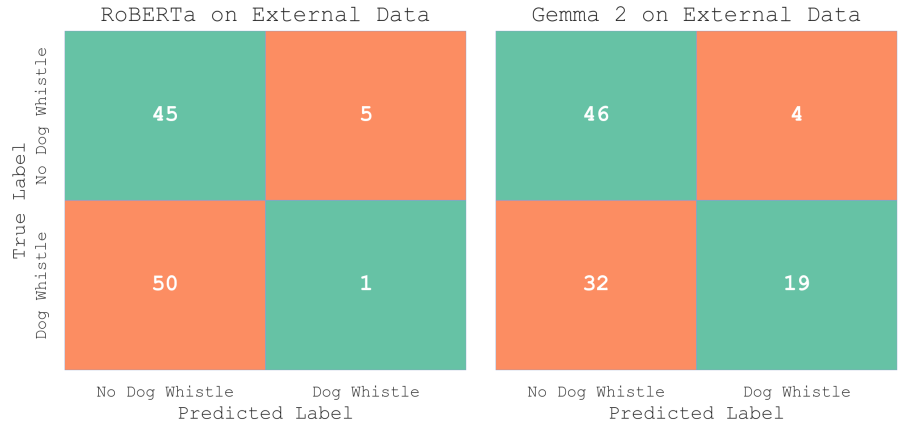

Figure 4.2: Confusion matrix for N = 101 external detection dataset for each finetuned model, showing the models' performance out of distribution.

The human baseline reported by Kruk et al. achieves an accuracy of 67%, which the fine-tuned Gemma 2 model nearly matches with 64%. However, when considering the F1 score for the dog whistle class, the limitations of Gemma 2 become evident. Despite its higher precision, the model struggles with recall (i.e., capturing most dog whistles), resulting in a lower F1 score of 0.51 compared to the human benchmark of 0.65. Compared to the GPT-4 model reported by Kruk et al., which achieves an accuracy of 85% and an F1 score of 0.86 in the dog whistle class, both fine-tuned models fall short. While Gemma 2 demonstrates moderate generalization with an F1 score of 0.51, RoBERTa performs significantly worse out of distribution, achieving an F1 score of only 0.04.

Table 4.3: Comparison classification metrics of fine-tuned models with GPT-4 and human baseline.

Metric	RoBERTa	Gemma 2	GPT-4*	Human Baseline*
F1-Score (Dog Whistle)	0.04	0.51	0.86	0.65
Accuracy	0.46	0.64	0.85	0.67

*Reported by Kruk et al. (2024).

4.3 Summary of the Results

In this chapter I discussed the performance of two fine-tuned models, RoBERTa and Gemma 2, on both in-distribution and out-of-distribution data. On the in-distribution validation set, RoBERTa outperformed Gemma 2, achieving higher accuracy and recall for the dog whistle class, and demonstrating superior overall classification performance. However, when evaluated on an external benchmark dataset, both models experienced a decline in performance, with RoBERTa showing a substantial drop in recall and F1 score for the dog whistle class. In contrast, Gemma 2 maintained more balanced performance across classes and demonstrated greater robustness under distribution shift. These results indicate that while RoBERTa is more effective within the training distribution, Gemma 2 is more reliable for detecting dog whistles in unseen, real-world contexts. Furthermore, comparisons with the human and GPT-4 baselines reported by Kruk et al. reveal a significant performance gap, particularly in the ability to handle the subtle and context-sensitive classification task of dog whistle detection.

Chapter 5

Discussion

The results presented above have produced three principal insights into automated dog-whistle detection. First, fine-tuned RoBERTa achieves promising in-distribution performance but fails to generalize under domain shift. Second, the bigger Gemma 2 model, while slightly weaker in-distribution, proves more robust on external data. Third, both models remain substantially behind human annotators and proprietary systems such as GPT-4, highlighting the difficulty of context-dependent classification tasks.

5.1 Differences Between Models and Distributions.

The differences in performance can be explained by the underlying architecture of the models. RoBERTa has the advantage its masked language modeling approach that could have driven the better in-distribution results, but at the same time caused overfitting to noisy data, leading to the drop in performance for the 101 external test cases. In contrast, the left to right embeddings in Gemma 2 might have reduced its performance in the in-distribution set, but the superior pretrained knowledge base of the bigger Gemma architecture led to more robust results in the external test set. Still, the drop in performance from in-distribution to external in both models shows that the problem does not necessarily lie in the model choice but in the underlying training data.

Two main issues can be identified between the in-distribution data and the external test data. Firstly, the training and validation data was in large parts not manually labeled, but through an automated process as explained in Section 3.2. Through the disambiguation method with GPT-4, according to Kruk et al. (2024), the LLM got a precision score of 0.92. While the exact accuracy is not reported by the authors, we can use the accuracy from the zero shot classification task to extrapolate the noisiness of the data at hand. Here, GPT-4 got 0.85 accuracy on the 101 test cases. This suggests that a non-negligible portion of the training data may contain mislabeled examples, introducing noise into the fine-tuning process. While this issue was partially mitigated by a manual verification of 2,000 null-class examples and removal of the remaining potential mislabels, inconsistencies likely remained, particularly within the positive class.

When looking at the confidence score of the confident learning label check from Section 3.2 in Figure 5.1, we can see that the median label score for instances labeled as positive is substantially lower than for those labeled negative. Importantly, this finding is counterintuitive. Given that the dataset construction by Kruk et al. was optimized for high precision in identifying positive cases, and only positive examples were publicly released, one might expect that the negative class, would contain more noise. However, the CL-derived label scores suggest the opposite: that the remaining positive instances are less confidently labeled on average, pointing

to a higher degree of uncertainty or inconsistency in this class. Furthermore, looking again at the temporal distribution of the training data in Figure 3.3, we observe a heavy temporal bias concentrated between 2016 and 2020. During this period, the dataset includes a significantly higher volume of potential dog whistle instances, likely reflecting the heightened political and cultural polarization of that time. As a result, many of the keywords frequently appearing in the data, such as "MAGA", are closely tied to that specific sociopolitical context. While these terms were highly salient during the dataset's peak period, they may now be outdated or less commonly used in current discourse.

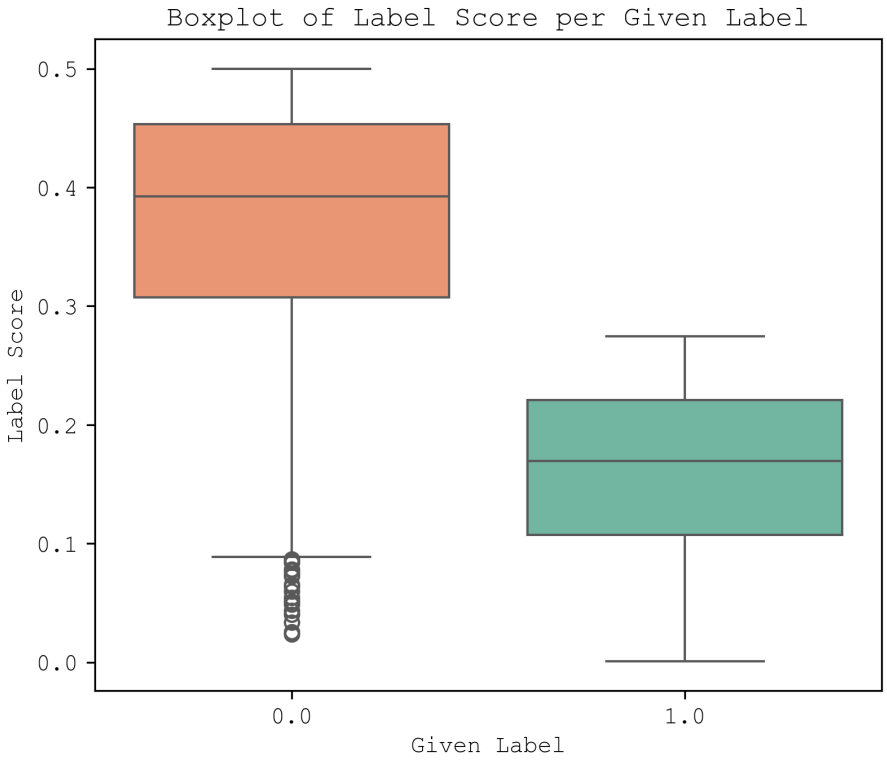

Figure 5.1: Box plot confidence score in CL pipeline for 6525 potential mislabels, representing the confidence of the model in the given label.

Secondly, while the external dataset of 101 manually reviewed instances provides a useful benchmark for evaluating model generalization, it remains relatively small in size. As such, caution is warranted when drawing strong conclusions about real-world performance based solely on this limited sample. Moreover, during manual inspection of the external test cases, three instances were identified as likely mislabeled [1]. In each of these cases, the labeling appeared inconsistent with the stated annotation guidelines or context. When correcting for these apparent labeling errors, the accuracy of the fine-tuned Gemma 2 model rises from 64% to 67%, matching the reported human baseline. This suggests that the observed performance gap between Gemma 2 and human annotators may be partially attributed to annotation noise in the benchmark itself, rather than an inherent shortcoming of the model. While this does not invalidate the general trend of performance decline under distribution shift, it points to the importance of both high-quality gold-standard test data and transparency in labeling criteria, particularly for tasks as semantically nuanced as dog whistle detection.

[1] See Appendix A.

5.2 Limitations

While this project demonstrates promising directions for the detection of dog whistles using fine-tuned language models, several limitations must be acknowledged.
First, the quality of the training data remains a central concern. Although efforts were made to clean and validate the dataset, including the use of GPT-4 for disambiguation and a confident learning pipeline for label auditing, the resulting labels are still subject to noise. As discussed in Section 3.2, much of the dataset was generated through automated processes rather than manual annotation. This introduces uncertainty into the model's learning signal, particularly for subtle and ambiguous cases, and likely impacts both precision and generalization.
Second, while the Gemma 2 model showed more robust out-of-distribution performance, newer generations of this model are availabe[2]. The same observation can be stated about RoBERTa: Warner et al. (2024) recently published an improved BERT variant called "modernBERT", which beats RoBERTa on many benchmarks. Future work could explore more modern small language models or instruction-tuned variants that offer better performance-efficiency trade-offs.
Third, the concept of a "dog whistle" itself poses challenges for computational modeling. The definitions used in this project, while based on prior work, are not always well-aligned with how such terms function in dynamic online discourse, as discussed in Section 2.2. Next to the differences in contexts, many dog whistles rely heavily on cultural or subcultural context, shifting over time and varying across platforms. As such, even well-trained models may struggle to keep pace with evolving language. The underlying training data is probably already outdated and needs to be updated periodically with new emerging dog whistles.
Fourth, it is important to note that this thesis focuses exclusively on dog whistles within an American sociopolitical context. The examples used, as well as the models' learned representations, are grounded in U.S.-centric discourse and may not generalize to other cultural or linguistic settings. Dog whistles in other countries may rely on different historical references, slang, or ideological codes, and would likely require region-specific datasets and annotation frameworks.
Finally, the task of dog whistle detection is inherently difficult, even for humans. Dog whistles are designed to be interpretable only by a targeted audience, making them intentionally ambiguous or covert. This characteristic challenges not only automated systems but also human annotators, and it underscores the importance of careful dataset construction, annotation transparency, and interdisciplinary approaches when addressing such nuanced linguistic phenomena.

[2]See: "huggingface.co/google/gemma-3-4b-it". Note, this model is twice as big as the Gemma-2-2b model used in this thesis, requiring more recourse intensive fine-tuning.

Chapter 6

Conclusions and Future Work

This thesis set out to explore the automated detection of dog whistles using fine-tuned language models, with a focus on their application in both political and online discourse. The results demonstrate that while promising, this task remains highly challenging due to the subtle, context-dependent nature of dog whistles and the limitations inherent in the data and tools currently available.

Three core findings emerged from the evaluation. First, the RoBERTa model achieved strong in-distribution performance but showed significant weaknesses when applied to out-of-distribution data. Second, the larger Gemma 2 model, though slightly weaker on in-distribution validation, generalized better to the external test set, suggesting more robust semantic representations due to its broader pretraining. Third, both models fall short of the performance achieved by human annotators and proprietary systems such as GPT-4, reaffirming the difficulty of dog whistle detection as a classification task.

The discussion revealed that model performance was not solely a function of architecture, but also deeply tied to the quality and scope of the training data. A key issue was the presence of label noise, stemming from the use of automated GPT-4-based labeling and a lack of comprehensive manual verification. Confidence analysis using the CL pipeline further suggested that positive instances were more likely to be misclassified, challenging the assumption that the dataset's focus on high-precision positive labeling would yield clean training signals. In addition, the temporal concentration of the data between 2016 and 2020 likely introduced bias toward political language specific to that era, potentially limiting model performance on more recent discourse.

The external benchmark used to assess generalization, while useful, was limited in size and itself not immune to labeling inconsistencies. Upon manual inspection, several instances were found to be potentially mislabeled. When corrected, the performance of the Gemma 2 model reached the human baseline, underscoring the impact of annotation quality on evaluation outcomes.

Despite these challenges, the study contributes meaningful insights to the field. It highlights both the potential and the limitations of large language models for detecting subtle, coded forms of speech that are increasingly relevant in digital political communication. Moreover, it underscores the importance of continual dataset refinement, transparent annotation procedures, and context-sensitive modeling when tackling such linguistically and ethically complex problems.

Future work should consider the adoption of more recent and efficient model architectures, such as instruction-tuned or multilingual variants, and expand the scope beyond the American context to capture the global evolution of coded political language. As online platforms continue to play a central role in shaping public discourse, the development of accurate and

adaptable systems for dog whistle detection will be critical not only for research but also for effective and equitable content moderation.

References

Albertson, B. L. (2015), 'Dog-Whistle Politics: Multivocal Communication and Religious Appeals', *Political Behavior* **37**(1), 3–26.
URL: *https://doi.org/10.1007/s11109-013-9265-x*

Angluin, D. and Laird, P. (1988), 'Learning from noisy examples', *Machine Learning* **2**(4), 343–370.
URL: *https://doi.org/10.1007/BF00116829*

Boholm, M. and Sayeed, A. (2023), Political dogwhistles and community divergence in semantic change, *in* N. Tahmasebi, S. Montariol, H. Dubossarsky, A. Kutuzov, S. Hengchen, D. Alfter, F. Periti and P. Cassotti, eds, 'Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change', Association for Computational Linguistics, Singapore, pp. 53–65.
URL: *https://aclanthology.org/2023.lchange-1.6*

Caselli, T., Basile, V., Mitrović, J. and Granitzer, M. (2021), HateBERT: Retraining BERT for Abusive Language Detection in English, *in* A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran and Z. Waseem, eds, 'Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)', Association for Computational Linguistics, Online, pp. 17–25.
URL: *https://aclanthology.org/2021.woah-1.3/*

Davidson, T. and Chae, Y. Y. (2025), 'Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning'.
URL: *https://osf.io/sthwk*$_v$2

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'.
URL: *http://arxiv.org/abs/1810.04805*

Drakulich, K., Wozniak, K. H., Hagan, J. and Johnson, D. (2020), 'Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics', *Criminology* **58**(2), 370–402.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9125.12239*

Gu, Y., Dong, L., Wei, F. and Huang, M. (2024), 'MiniLLM: Knowledge Distillation of Large Language Models'.
URL: *http://arxiv.org/abs/2306.08543*

Haney-López, I. (2014), *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*, Oxford University Press.

Hartmann, D., Oueslati, A., Staufer, D., Pohlmann, L., Munzert, S. and Heuer, H. (2025), 'Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations'.
URL: *http://arxiv.org/abs/2503.01623*

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D. and Kamar, E. (2022), ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection, arXiv. Version Number: 4.
URL: *https://arxiv.org/abs/2203.09509*

Henderson, R. and McCready, E. (2018), How Dogwhistles Work, *in* S. Arai, K. Kojima, K. Mineshima, D. Bekki, K. Satoh and Y. Ohta, eds, 'New Frontiers in Artificial Intelligence', Springer International Publishing, Cham, pp. 231–240.

Hertzberg, N., Cooper, R., Lindgren, E., Rönnerstrand, B., Rettenegger, G., Breitholtz, E. and Sayeed, A. (2022), 'Distributional properties of political dogwhistle representations in Swedish BERT', *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* pp. 170–175. Conference Name: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) Place: Seattle, Washington (Hybrid) Publisher: Association for Computational Linguistics.
URL: *https://aclanthology.org/2022.woah-1.16*

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2021), 'LoRA: Low-Rank Adaptation of Large Language Models'.
URL: *http://arxiv.org/abs/2106.09685*

Iyengar, S., Sood, G. and Lelkes, Y. (2012), 'Affect, Not Ideology: A Social Identity Perspective on Polarization', *Public Opinion Quarterly* **76**(3), 405–431.
URL: *https://doi.org/10.1093/poq/nfs038*

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F. and Rolnick, D. (2022), 'Aligning artificial intelligence with climate change mitigation', *Nature Climate Change* **12**(6), 518–527. Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41558-022-01377-7*

Khoo, J. (2017), 'Code Words in Political Discourse', *Philosophical Topics* **45**(2), 33–64. Publisher: University of Arkansas Press.
URL: *https://www.jstor.org/stable/26529437*

Kruk, J., Marchini, M., Magu, R., Ziems, C., Muchlinski, D. and Yang, D. (2024), 'Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles'.
URL: *http://arxiv.org/abs/2406.06840*

Li, C. and Flanigan, J. (2023), 'Task Contamination: Language Models May Not Be Few-Shot Anymore'.
URL: *http://arxiv.org/abs/2312.16337*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'.
URL: *http://arxiv.org/abs/1907.11692*

Magu, R., Joshi, K. and Luo, J. (2017), 'Detecting the Hate Code on Social Media'.
URL: *http://arxiv.org/abs/1703.05443*

Mendelberg, T. (2001), *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*, Princeton University Press.
URL: *https://www.jstor.org/stable/j.ctt1trkhws*

Mendelsohn, J., Bras, R. L., Choi, Y. and Sap, M. (2023), 'From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models'.
URL: *http://arxiv.org/abs/2305.17174*

Mo, K., Liu, W., Xu, X., Yu, C., Zou, Y. and Xia, F. (2024), Fine-Tuning Gemma-7B for Enhanced Sentiment Analysis of Financial News Headlines, *in* '2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)', pp. 130–135.
URL: *https://ieeexplore.ieee.org/abstract/document/10594605*

Morin, R. (1988), 'Opinion | BEHIND THE NUMBERS CONFESSIONS OF A POLLSTER', *The Washington Post* .
URL: *https://www.washingtonpost.com/archive/opinions/1988/10/16/behind-the-numbers-confessions-of-a-pollster/3523c065-11b5-42ba-9986-c317bdecf2dd/*

Mozafari, M., Farahbakhsh, R. and Crespi, N. (2020), A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media, *in* H. Cherifi, S. Gaito, J. F. Mendes, E. Moro and L. M. Rocha, eds, 'Complex Networks and Their Applications VIII', Springer International Publishing, Cham, pp. 928–940.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A. (2024), 'A Comprehensive Overview of Large Language Models'.
URL: *http://arxiv.org/abs/2307.06435*

Nguyen, T. T., Wilson, C. and Dalins, J. (2023), 'Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts'.
URL: *http://arxiv.org/abs/2308.14683*

Northcutt, C. G., Athalye, A. and Mueller, J. (2021), 'Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks'.
URL: *http://arxiv.org/abs/2103.14749*

Ocampo, N. B., Sviridova, E., Cabrio, E. and Villata, S. (2023), An In-depth Analysis of Implicit and Subtle Hate Speech Messages, *in* A. Vlachos and I. Augenstein, eds, 'Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Dubrovnik, Croatia, pp. 1997–2013.
URL: *https://aclanthology.org/2023.eacl-main.147*

Safire, W. (2008), *Safire's Political Dictionary*, Oxford University Press.

Saleh, H., , Areej, A., and Moria, K. (2023), 'Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model', *Applied Artificial Intelligence* **37**(1), 2166719.
URL: *https://doi.org/10.1080/08839514.2023.2166719*

Sayeed, A., Breitholtz, E., Cooper, R., Lindgren, E., Rettenegger, G. and Rönnerstrand, B. (2024), 'The utility of (political) dogwhistles – a life cycle perspective'. Publisher: John Benjamins.
URL: *https://www.jbe-platform.com/content/journals/10.1075/jlp.23047.say*

Sen, T., Das, A. and Sen, M. (2024), 'HateTinyLLM : Hate Speech Detection Using Tiny Large Language Models'.
URL: *http://arxiv.org/abs/2405.01577*

Stanley, J. (2015), *How Propaganda Works*, Princeton University Press.
URL: *https://www.jstor.org/stable/j.ctvc773mm*

Suhay, E., Bello-Pardo, E. and Maurer, B. (2018), 'The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments', *The International Journal of Press/Politics* **23**(1), 95–115. Publisher: SAGE Publications Inc.
URL: *https://doi.org/10.1177/1940161217740697*

Team Gemma, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., Amersfoort, J. v., Gordon, J., Lipschultz, J., Newlan, J., Ji, J.-y., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R. and Andreev, A. (2024), 'Gemma 2: Improving Open Language Models at a Practical Size'.
URL: *http://arxiv.org/abs/2408.00118*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u. and Polosukhin, I. (2017), Attention is All you Need, *in* 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html*

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J. and Poli, I.

(2024), 'Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference'.
URL: *http://arxiv.org/abs/2412.13663*

Wetts, R. and Willer, R. (2019), 'Who Is Called by the Dog Whistle? Experimental Evidence That Racial Resentment and Political Ideology Condition Responses to Racially Encoded Messages', *Socius* **5**, 2378023119866268. Publisher: SAGE Publications.
URL: *https://doi.org/10.1177/2378023119866268*

Witten, K. (2023), 'The Definition and Typological Model of a Dogwhistle', *Manuscrito* **46**, e. Publisher: UNICAMP - Universidade Estadual de Campinas, Centro de Lógica, Epistemologia e História da Ciência.
URL: *https://www.scielo.br/j/man/a/vktmMPLqW4PSWHJhHKmLCfh/*

Appendix A

Supporting Tables and Figures

Prompt for the Gemma Model

"You are a helpful assistant participating in a content moderation task of dog whistle detection. A dog whistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. Does the following sentence contain a dog whistle? [Reddit Post]". This prompt was adapted from Kruk et al. (2024).

Data Set Description and LoRA modules

Table A.1: Target Modules for LoRA Adaptation of Gemma 2. Injecting in both the classification head (MLP) and pretrained Model.

Module Name	Description
`q_proj`	Query projection in the attention mechanism
`k_proj`	Key projection in the attention mechanism
`v_proj`	Value projection in the attention mechanism
`o_proj`	Output projection in the attention mechanism
`gate_proj`	Gate projection in the MLP block
`up_proj`	Up projection in the MLP block
`down_proj`	Down projection in the MLP block

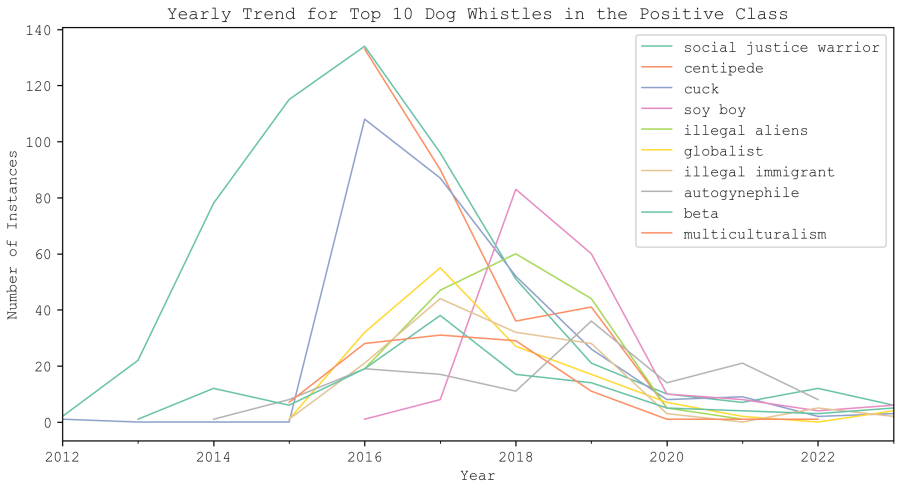

Figure A.1: Yearly frequency trends of the top 10 dog whistle terms in the positive class between 2012 and 2023. The terms "social justice warrior", "centipede", and "cuck" peaked sharply around 2016–2017, coinciding with heightened political polarization in the U.S. during the Trump election cycle. Other terms like "soy boy", "illegal aliens", and "globalist" also showed notable activity during this period before declining in frequency. Highlighting the bias in the distribution towards the years 2016 to 2020.

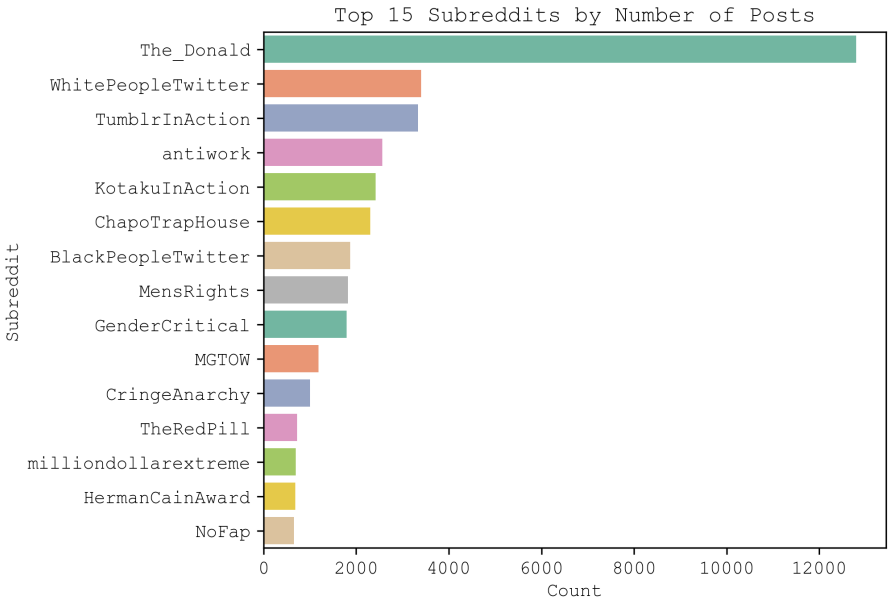

Figure A.2: Top 15 Subreddits by total number of posts in the dataset. The subreddit The_Donald overwhelmingly dominates in post count, contributing more than three times the volume of the next highest Subreddit. Other notable contributors include WhitePeopleTwitter, TumblrInAction, and antiwork, indicating a wide range of ideological and cultural communities represented in the dataset.

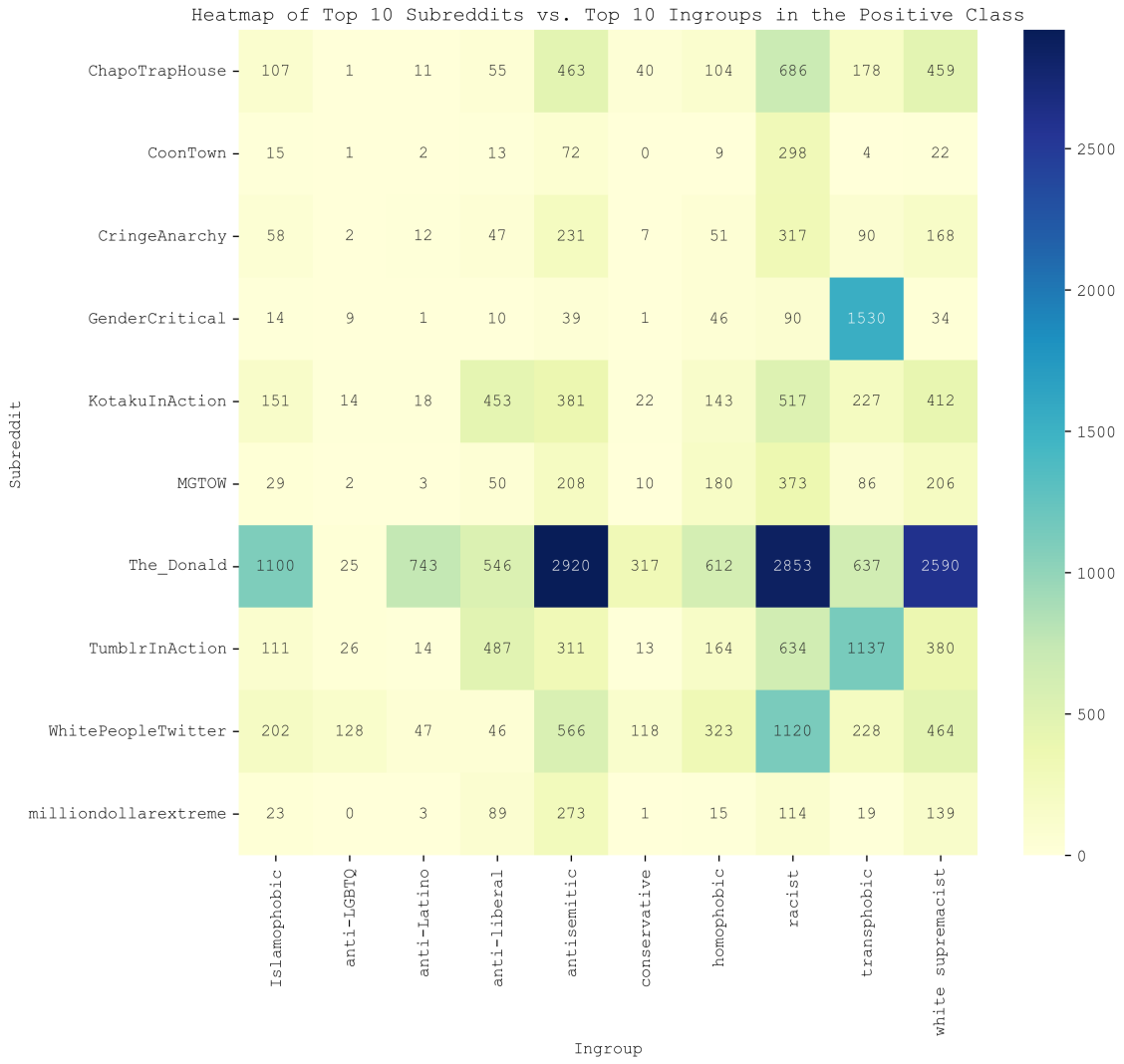

Figure A.3: Heatmap showing the distribution of the top 10 subreddits in the positive class across the top 10 targeted ingroups. The color intensity represents the number of dog whistle instances identified per Subreddit–ingroup pair. The Subreddit The_Donald stands out with consistently high frequencies across multiple ingroups, particularly in antisemitic, racist, and white supremacist categories. Other subreddits such as WhitePeopleTwitter and TumblrInAction also show high counts, particularly in the racist and transphobic dimensions.

Appendix B

Declaration of Authorship

I, Lino Zurmühl, I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on 28.04.2025 is identical to the printed version I submitted to the Examination Office on 29.04.2025.

In accordance with the Hertie School's AI Guidelines (February 2023), I declare that I have used Gemini (Latex formatting), Copilot (code debugging) and DeepL (grammar checks) in accordance with the Hertie School's policy on academic integrity.
I give consent for my work to be made available more widely to members of the Hertie School and the public with interest in teaching, learning and research.

Lino Zurmühl
April 29, 2025