Making FETCH! Happen: Finding Emergent Dog Whistles Through Common Habitats

Kuleen Sasse[1], Carlos Aguirre[1], Isabel Cachola[1],
Sharon Levy[2], Mark Dredze[1]
[1]Johns Hopkins University, [2]Rutgers University
Correspondence: ksasse1@jh.edu

Abstract

WARNING: This paper contains content that maybe upsetting or offensive to some readers. Dog whistles are coded expressions with dual meanings: one intended for the general public (outgroup) and another that conveys a specific message to an intended audience (ingroup). Often, these expressions are used to convey controversial political opinions while maintaining plausible deniability and slip by content moderation filters. Identification of dog whistles relies on curated lexicons, which have trouble keeping up to date. We introduce **FETCH!**, a task for finding emergent dog whistles in massive social media corpora. We find that state-of-the-art systems fail to achieve meaningful results across three distinct social media case studies. We present **EarShot**, a strong baseline system that combines the strengths of vector databases and Large Language Models (LLMs) to efficiently and effectively identify new dog whistles. [1]

1 Introduction

> *"white house petition: end **dual citizens** from serving in us and state governments"*
>
> — **Anonymous Gab User**

When a person writes **dual citizens** it may seem like an innocuous phrase, but many understand this as an antisemitic reference. The use of the phrase builds on centuries old questions of the loyalty of Jewish people towards governments, which recently has become a claim that all American Jews are double citizens of Israel and the United States.

This phrase is a dog whistle, which is defined as an "expression that sends one message to an outgroup while at the same time sending a second (often taboo, controversial, or inflammatory) message to an ingroup (Henderson and McCready,
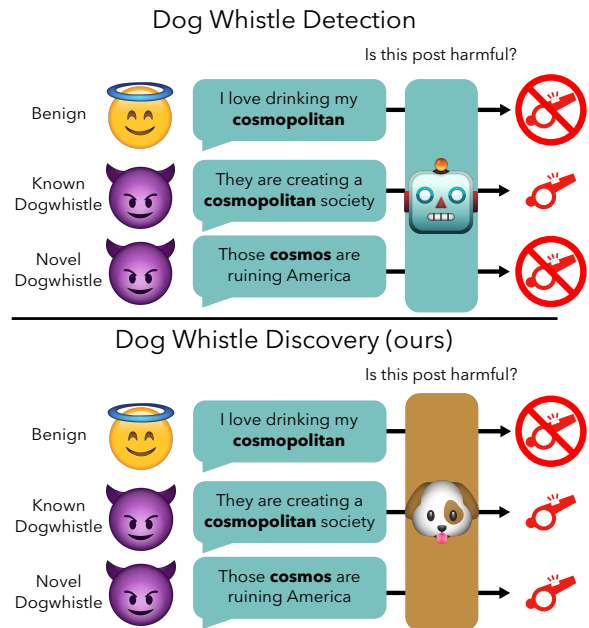

Figure 1: Comparison between dog whistle detection and our method of dog whistle discovery

2018)." Dog whistles are a common mainstay in today's turbulent political climate. For example, they prominently featured both on social media and in person in the 2016 United States presidential election (Tilley et al., 2020; Drakulich et al., 2020).

As dog whistles are a common way to transmit hate speech to a wide audience both digitally and offline, researchers have started to develop methods to detect them. However, dog whistles are notoriously difficult to find as they often slip by content moderation, toxicity, and hate speech filters at a high rate because of their benign second meaning (Mendelsohn et al., 2023). Going back our example, dual citizenship isn't an issue, but "dual citizen" in the above context has a specific meaning. For more examples of dog whistles and similar explanations of their meaning and difficulty are provided in Appendix B.

---

[1]Code for analysis and evaluation can be found at https://github.com/KuleenS/FETCH-Dog-Whistle

Currently, detecting dog whistles relies on manually curated lexicons, which can be effective but are labor-intensive to create and maintain. They cannot adapt to the dynamic evolution of language. For example (Figure 1), current methods can distinguish between the two meanings of the word "cosmopolitan": its use as the popular mixed drink and its use as a dog whistle to denigrate a multicultural and intellectual society. However, they miss a hypothetical linguistic shift of the term "cosmopolitan" to the shortening of "cosmos."

We consider the task of dog whistle discovery, a novel task where systems identify emerging or previously unknown dog whistles. To support this effort, we create **FETCH!** or **F**inding **E**mergent Dog Whistles **T**hrough **C**ommon **H**abitats, a benchmark designed to evaluate models' ability to detect emerging or novel dog whistles across diverse settings. Using the **FETCH!** benchmark, we evaluate three state of the art methods across three different social media case studies. Our evaluation reveals that these methods perform poorly, achieving an F-score of less than 5% across all case studies.

We introduce our own method as a strong baseline: **EarShot**. This method finds new dog whistles by leveraging the power of sentence embeddings and Large Language Models (LLMs). EarShot demonstrates significant improvements, with F-score increases ranging from 2 to 20 points across the social media case studies.

Summary of contributions

- Development of **FETCH!** benchmark for dog whistle discovery

- Rigorous evaluation of multiple state-of-the-art methods on **FETCH!**

- Creation of **EarShot**, a strong baseline architecture achieving SoTA performance on **FETCH!**

2 Related Work

2.1 Implicit Hate Speech Detection

Detecting novel dog whistles can be thought of as a form of hate speech detection, specifically implicit hate speech detection. Implicit hate speech can be defined as hate speech that uses coded or indirect language (Waseem et al., 2017). Many studies have focused on this task through creating new datasets (Hartvigsen et al., 2022; ElSherief et al., 2021; Ocampo et al., 2023; Breitfeller et al., 2019),

novel developing detection methods (Zhang et al., 2024; Gao et al., 2017), and methods to explain the hate speech (Qian et al., 2019). However, these methods focus on the overall text or post rather than the specific words and phrases.

2.2 Dog Whistles

Linguistics, political science, and other social sciences have researched the use of dog whistles and their appearance through various forms of analysis. Linguistic methods have focused on the semantics and pragmatics of their use (Henderson and McCready, 2018; Torices, 2021; Khoo, 2017; Saul, 2018; Bhat and Klein, 2020; Quaranto, 2022) and have further extended their work using agent-based simulations of the emergence of new dog whistles (Henderson and McCready, 2020). Historians have focused on showing the evolution and propagation of racial and dog whistle speech for political purposes and their lasting effects on society (Mendelberg, 2001; Haney-López, 2014). Sociologists and psychologists have focused on analyzing the attitudes of people before and after reading messages with implicit racial appeals, such as dog whistles (Wetts and Willer, 2019; Albertson, 2015). Other sociological work has studied the development and changes of dog whistles over time on the Web (Åkerlund, 2022). Political science has investigated the use of dog whistles and how they change public opinions and policies surrounding controversial topics such as sanctuary cities and other criminal justice policies (Lasch, 2016; Hurwitz and Peffley, 2005; Goodin and Saward, 2005).

2.3 Euphemism Detection

Euphemism detection focuses on finding novel euphemisms, terms, or phrases used to substitute more offensive terms to downplay its unpleasantness, a semantic phenomenon similar to dog whistles. Typically, the task is structured as a user is given a set of seed euphemisms, and the goal of the task is to find other euphemisms in a provided corpora, a process similar to lexicon induction (Lee et al., 2022).

Early work in euphemism detection often focused on using Word2Vec (Mikolov et al., 2013) to find euphemisms. In one work, the authors trained a Word2Vec model on the corpus of interest, and then, they used the similarities between the word vectors to obtain similar words to their seed words (Magu and Luo, 2018). Other works improved on this method by combining both Word2Vec and

dependency based word embeddings (Levy and Goldberg, 2014; Taylor et al., 2017), using two different Word2Vec models on two different datasets and finding the words with the greatest difference in embeddings (HADA et al., 2020; Yuan et al., 2018; Zhao et al., 2016), or using the difference between the two Word2Vec learned matrices (Aoki et al., 2017). Other works used methods combining Word2Vec methods with additional user and search behaviors of users (Yang et al., 2017). Later works further refined these methods by using both more powerful Masked Language Models like BERT (Devlin et al., 2019; Zhu et al., 2021) and Span-BERT (Joshi et al., 2020; Zhu and Bhat, 2021) both in a zeroshot setting (Maimaitituoheti et al., 2022) and a supervised setting and Large Language Models like GPT-3 (Brown et al., 2020) in a zeroshot setting (Keh, 2022).

2.4 Dog Whistle Detection

Relatively few studies have addressed the detection of dog whistles. Among these, three notable works focus specifically on Swedish dog whistles, employing both embedding-based methods (Boholm et al., 2024; Hertzberg et al., 2022) and analyzing semantic shifts (Boholm and Sayeed, 2023) In English, only two studies have focused on detecting dog whistles. In the first paper, they curated a set of English dog whistles to assess both large language models' (LLMs) ability to surface these terms and the impact on hate speech detectors. However, they only perform a small analysis with a handful of positive examples to see if LLMs can detect novel dog whistles (Mendelsohn et al., 2023). In the second paper, they introduced the largest dataset of posts containing dog whistles, filtered from sources such as Reddit and the Congressional Record by LLMs. In addition to the creation of the dataset, they conducted a small preliminary test of LLMs' capabilities in detecting novel dog whistles (Kruk et al., 2024). However, this test was curated to be balanced in the number of posts with and without dog whistles.

3 Task Description

We propose a novel task: **Finding Emergent Dog Whistles in Common Habitats** or **FETCH!** based on a similar task in previous euphemism detection papers (Joshi et al., 2020; Zhu and Bhat, 2021). For the task, a system is provided with a corpus and a set of initial seed dog whistles. The corpus might

be social media posts, forums, website pages, etc. The initial set of seed dog whistles is a list of known dog whistles that occur within the corpus. The goal of the task is to use the initial seed words and the corpus to discover other known dog whistles called the emergent dog whistles. These dog whistles are semantically similar to the seed dog whistles as they are used in similar scenarios but have different users, uses, and definitions. However, these dog whistles are not fully novel but they are unknown to the system or model. The goal is to measure the ability of the model or system to extrapolate from the seed dog whistles, hopefully leveraging the shared linguistic properties.

3.1 Case Studies

We describe the three example case studies of the FETCH! task that represent real-world or synthetic scenarios.

3.1.1 Synthetic (Reddit)

Our first scenario is an idealized setting where every post contains an associated dog whistle. We use the Silent Signals dataset from a recent paper (Kruk et al., 2024), which includes 16,000 posts scraped from Reddit from 45 controversial subreddits. Each post in the dataset contains a dog whistle that was identified by GPT-4 (OpenAI et al., 2024).

3.1.2 Balanced (Gab)

The balanced scenario represents a context with a higher-than-average prevalence of dog whistles with large amount of posts. For this scenario, we chose to use data from Gab, an alt-right alternative to Twitter where dog whistles and other forms of hate speech are more frequent. The data was collected over a short period using a web scraper that sequentially downloaded post IDs yielding approximately 300,000 posts.

3.1.3 Realistic (Twitter)

The realistic scenario reflects a typical online environment where dog whistles are sparse, and most posts are benign. For this scenario, we selected X (formerly Twitter) as the platform, as it contains some hate speech but removes much of it through content moderation. Data were collected using the 1% public Twitter API stream between January 2017 and December 2022. Given the dataset's massive size (approximately 1 billion tweets), we subsampled it to make it manageable. We randomly sampled 6 million English language tweets

without dog whistles and 1 million tweets containing possible dog whistles identified using regular expressions from a set of ground truth dog whistles. This process resulted in a dataset of 7 million tweets.

3.2 Seed Dog Whistles

We use the dog whistle dataset released by Mendelsohn et al. (2023) as our ground truth set of dog whistles for our Balanced and Realistic scenarios as the Synthetic scenario has its own set of dog whistles. The dataset is a hand-curated taxonomy of around 340 root dog whistles from community wikis and academic sources. Each root dog whistle comes other surface forms of the root dog whistle. Table 1 lists a summary of the number of dog whistles found in each scenario. The full list appears in Appendix A.

Scenario	Balanced	Synthetic	Realistic
Total Dog Whistles Found	77	298	177
Reference Dog Whistles Total	340	298	340

Table 1: Number of dog whistles root forms found in each scenario

For each corpora, we divide the dog whistles that are able to be found in each scenario into train/test splits (20% / 80%), creating a set of seed dog whistles to use to find the rest of the dog whistles. We split the dog whistles using stratified sampling based on the length of phrase in n-grams. We do a stratified split so as not to unfairly punish methods that can identify larger n-grams. After splitting, we add the surface forms as well to each split in order to cover all variants of the dog whistles. We split on n-grams instead of on different metadata provided in the dog whistle ground truth dataset like the group being targeted or the user of the dog whistle as the Synthetic scenario does not provide this information and it is more realistic scenario to have a set of randomly chosen dog whistles rather a well distributed set of dog whistles according to some metadata.

3.3 Labeling Process

To label the possible dog whistles in each post for the Balanced and Realistic Scenarios, we opted to use regular expressions to find them quickly and cheaply in both scenarios. We opted for regular expressions over human annotations, LLMs, or training our own classifier on human annotations as dog whistles are fairly unique and do not result

in many false positives. In addition, these methods would require significant computational costs that are infeasible for our systems.

3.4 Metrics for Evaluation

We use precision, recall and F-score to measure the performance of different methods for our task. We opt to use precision, recall, and F-score without a threshold as it is a more general metric as some models or systems do not rank their predictions. In addition, we do not create metrics that normalize for the maximum length of the words output as in testing, models with smaller length outputs have inflated scores.

To deal with the problem of dog whistles having multiple surface forms, all metrics count a positive match for a single root dog whistle as returning an exact match of at least one of the surface forms or the root itself.

Precision is computed as normal, but we modify recall to normalize by only the positive terms that exist in the corpus (Data Potential Recall, DPR). We report $F_{0.5}$ scores using DPR and Precision instead of F_1 to prefer methods with a high precision to reduce the load on a human reviewer.

4 Experimental Setup

4.1 Models

We test four different methods on **FETCH!**: Word2Vec/Phrase2Vec (Mikolov et al., 2013), MLM (Zhu et al., 2021), EPD (Zhu and Bhat, 2021), and our method: EarShot.

4.1.1 Word2Vec/Phrase2Vec

Word2Vec and Phrase2Vec are widely used models for euphemism detection due to their lightweight nature and fast computation on large datasets. These models capture semantic relationships between words or phrases, enabling the detection of euphemistic language. In our implementation of these models for dog whistle discovery, we adopt a methodology similar to (Magu and Luo, 2018).

To create the training data for the Word2Vec/Phrase2Vec (Mikolov et al., 2013), we preprocessed the corpus by removing all stop words and lemmatizing the rest using littlebird (DeLucia, 2020). After this preprocessing step, we replaced URLs, hashtags, retweets, and mentions with special symbols like @USER and HTTPURL. We also replaced emojis with their official English aliases using the emoji package (Kim and Wurster,

2024). Finally, we tokenized the result with the NLTK Tweet Tokenizer (Bird and Loper, 2004).

To construct the Phrase2Vec models, we utilized the Gensim Phraser module (Rehurek and Sojka, 2011) to merge common n-gram phrases in the original corpus, generating 2-gram and 3-gram versions of the models. We only train up to 3-gram models as a majority of dog whistles in our dataset are 3-grams and smaller. We then trained models using Gensim with a maximum vocabulary size of 500,000 and a context window of 5 with a vector dimension of 100 for 10 epochs.

To generate the predictions, we provided the Word2Vec/Phrase2Vec model the list of seed dog whistles and retrieved the top 10 most similar words to each seed word. Then, we queried for the top 10 most similar words to those words. We continued this process 10 times creating 10 levels of predictions.

4.1.2 MLM

Euphemism detection relies heavily on context, but static embeddings from Word2Vec/Phrase2Vec fail to capture this. To address this, researchers leveraged contextual embeddings from Masked Language Models like BERT (Zhu et al., 2021). The MLM system first samples 2000 sentences from the input dataset containing the seed dog whistles. We sampled 2000 sentences as it is the default parameter provided. For each sampled sentence, it masks out the seed dog whistle in that sentence and uses the Masked Language Model to fill in the mask by creating a distribution over possible replacements and then returning the top-k most likely tokens. We test this method using two MLM encoder models: Bernice (DeLucia et al., 2022) and BERTweet (Nguyen et al., 2020). Both these models have been adapted to social media data, providing a reasonable benchmark as pretraining from scratch or finetuning custom models is computationally expensive as corpora grow to social media scale. We test the method at thresholds for k of $\{50, 100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600\}$.

4.1.3 EPD

The Euphemistic Phrase Detector (EPD) (Zhu and Bhat, 2021) builds upon the MLM model by predicting phrases instead of single tokens. The EPD model uses Autophrase (Shang et al., 2017), a statistical phrase extractor to extract important phrases from the corpus. Next, it trains a uni-

gram Word2Vec model (Mikolov et al., 2013) on the same corpus to filter down the phrases by selecting the top 1000 most similar phrases from AutoPhrase to the seed dog whistles. Finally, it samples 2000 sentences from the input dataset that contain the seed dog whistles. For each sampled sentence, it replaces the seed dog whistle with each predicted phrase from the previous stage and uses the Masked Language Model to calculate the probability of the sentence with the replaced phrase. The sentences are sorted by their calculated probabilities, and the top-k sentences are selected and the corresponding inserted phrase for each sentence is returned as the predictions. We test this method using SpanBERT (Joshi et al., 2020) as our Masked Language Model as it is the default model used. We test the method at increasing thresholds for k of $\{50, 100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600\}$.

4.1.4 EarShot

LLMs represent the next step in advancing NLP capabilities. Using LLMs, we create EarShot, a strong baseline architecture for the task, structured in three stages.

In the first stage, all posts in the corpora are turned into vectors using a sentence transformer (Reimers and Gurevych, 2019) and put into a vector database. We opt to use the `all-MiniLM-L6-v2` as our sentence transformer as it is one of the fastest models with high quality embeddings. We use ChromaDB (Chroma, 2024) as our vector database.

In the second stage, we obtain the vectors for the posts containing the seed dog whistles and find the closest vector to each seed post vector that are not the seed post themselves. By identifying closet neighbors, we aim to capture posts that are likely related in meaning, sentiment, or intent, even if they do not share exact word matches. In addition, this reduces the amount of computational resources used to find dog whistles in the scenarios. After this step, we try two different methods to extract the predicted dog whistles: DIRECT and PREDICT, labeled (1) and (2) in Figure 2 respectfully.

For the DIRECT pipeline, we pass all the posts to a LLM to extract the dog whistles from the posts provided from the vector lookup. The model is prompted to list all dog whistles in a structured JSON format with a prompt from (Kruk et al., 2024). The prompt is shown in Table 7 in Appendix C. We evaluate three LLMs: `meta-llama/Meta-Llama-3.1-8B-Instruct`
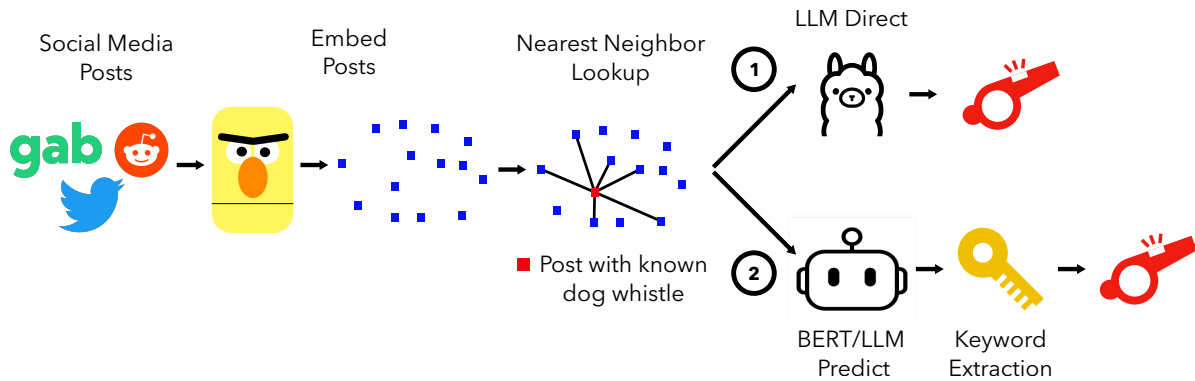
Figure 2: Flowchart for the EarShot System. Two paths can be taken after doing the nearest neighbor lookup. Path (1) is the more computationally expensive system that asks directly prompts a LLM for an answer. Path (2) is the cheaper method that leverages keyword extraction models and binary prediction.

(LLaMa 8B) (Dubey et al., 2024), `mistralai/Mistral-7B-Instruct-v0.3` (Mistral 7B) (Jiang et al., 2023), and `meta-llama/Llama-2-13b-chat-hf` (LLaMa 13B) (Touvron et al., 2023). We choose to test LLaMa and Mistral as they some of the best performing open source LLMs. We choose to use smaller open source models over larger closed source models as they are easier to reproduce, provide a strong baseline, and scale better to larger datasets.

For the PREDICT pipeline, we have two stages. In the first stage, we have a filtering step. In the second stage, we pass the posts to a keyword extractors and return the top-k predictions of them.

For the filtering stage, we filter the posts using either a BERT model or a LLM. For the BERT based filtering, we use an off the shelf hate speech or toxicity detection model to classify the posts. We remove all the posts that are labeled not toxic or hate speech according to the model as dog whistles are commonly found in hate speech. We call this method filtering BERT-PREDICT. We test three popular hate speech and toxicity classifiers: `tomh/toxigen_hatebert` (ToxiGen BERT) (Hartvigsen et al., 2022), `facebook/roberta-hate-speech-dynabench-r4-target` (RoBERTa R4) (Vidgen et al., 2021), and `Hate-speech-CNERG/bert-base-uncased-hatexplain` (HateXplain BERT)(Mathew et al., 2022).

For the LLM based filtering, we prompt the model to reply yes or no if there is a dog whistle in the post. We structure our prompt similarly to Silent Signals (Kruk et al., 2024), shown in Table 7 in Appendix C. We keep all the posts that the model explicitly generates a yes answer to using a substring search. We call this filtering method LLM-PREDICT. We test the same three models used in the DIRECT pipeline for similar reasons.

We opted to use off the shelf hate speech classifiers and zero shot LLMs as training our own classifiers performed poorly. To train our own, we finetuned off the shelf hate speech classifiers with the seed posts as positive and an equal number of posts randomly sampled from the rest of the corpus as negative posts. We sampled randomly negative posts because sampling posts without dog whistles would have been cheating. After finetuning these models for 2 epochs with 2e-5 learning rate and 0.01 weight decay. These models did not predict any posts have dog whistles and achieved 0 on all three metrics.

After the filtering step, we pass the posts to a keyword extractors and return the top-k predictions of them. We use five different popular keyword extraction algorithms: KeyBERT (Grootendorst, 2020), RAKE (Rose et al., 2010), YAKE, (Campos et al., 2018), TextRank (Mihalcea and Tarau, 2004), and TF-IDF (Spärck Jones, 1972). If the keyword extractor requires a specific range of n-grams to return, we test ranges of 1-gram, 1–2-grams, and 1–3-grams. We report the performance of the method at the same thresholds in MLM and EPD.

We filter posts before passing them to keyword extractors for two reasons. As keyword extractors are mostly topic agnostic, it is better to get similar hateful posts before getting the most important. Second, some keyword extractors do not scale well to large amounts of posts seen in social media corpora. For example, TextRank (Mihalcea and Tarau,

2004) breaks text into words and then converts the words into a graph with the edges being mathematical relations between the words. This can lead to a very large graph that is impossible to process with our current computational resources.

5 Results

5.1 Word2Vec/Phrase2Vec

Scenario	Model	Level	Prec	DPR	$F_{0.5}$
Balanced	Unigram	1	**1.64**	4.12	**1.86**
	Bigram	1	0.83	**6.17**	1.00
	Trigram	1	0.83	**6.17**	1.00
Synthetic	Unigram	1	**5.50**	8.40	**5.91**
	Bigram	2	2.82	**21.85**	3.42
	Trigram	1	3.24	7.98	3.67
Realistic	Unigram	1	**2.97**	1.47	**2.47**
	Bigram	1	0.89	1.47	0.97
	Trigram	1	1.03	1.47	1.10

Table 2: Results for best Word2Vec/Phrase2Vec models by $F_{0.5}$. Prec is Precision. DPR is Data Potential Recall. Best scores for across each dataset are **bolded**

We summarize the best models for Word2Vec and Phrase2Vec in Table 2. Unigram models do the best across all three datasets both in terms of Precision and $F_{0.5}$ but it comes at the cost of lower recall as they cannot capture more complex phrases.

On the Synthetic scenario, none of the models even hit 7-10 F-score, which means they are predicting most words wrong even though all sentences have a dog whistle. On Realistic and Balanced, performance significantly drops as models can only achieve 1-2 F-score showing how sparser environments hurt the Word2Vec model even more.

Most models perform best at the first or second prediction level. This conservative strategy limits the model's usefulness, as it predicts only a small number of new dog whistles. Threshold choice figures are in Appendix D. Overall, unigram Word2Vec models excel in this task but struggle to capture more complex phrases due to model limitations and their conservative prediction approach.

5.2 MLM and EPD

Table 3 presents the top-performing MLM and EPD models for each scenario. Overall, MLM and EPD models perform extremely poorly across all metrics. They obtain less than 1 point in F-score, very poor precision around 1-2 points, and a modest DPR around 1-9 points. For the MLM models, we attribute this lackluster performance to limitation

Scenario	Model	Threshold	Prec	DPR	$F_{0.5}$
Balanced	MLM Bernice	50	0.14	6.58	0.17
	MLM BERTweet	50	**0.23**	**9.47**	**0.28**
	EPD SpanBERT	50	0.00	0.00	0.00
Synthetic	MLM Bernice	400	0.50	0.84	0.54
	MLM BERTweet	50	**2.00**	0.42	1.14
	EPD SpanBERT	800	1.85	**3.36**	**2.03**
Realistic	MLM Bernice	6400	0.05	**2.21**	0.06
	MLM BERTweet	400	0.50	1.47	0.58
	EPD SpanBERT	200	**0.58**	0.74	**0.61**

Table 3: Results for best MLM and EPD models by $F_{0.5}$. Prec is Precision. DPR is Data Potential Recall. Best scores for across each dataset are **bolded**

of the tokenizer of the models as MLM models are restricted to predicting single tokens learned from their training corpora. As for the EPD model, we believe that the filtering step using a unigram Word2Vec model leaves out complex phrases.

Most of the models only predict 50 words mimicking failures in Word2Vec and Phrase2Vec. Additional figures detailing the choice of threshold are provided in Appendix E. Overall, MLM and EPD models lag behind the simpler embedding models.

5.3 EarShot

5.3.1 PREDICT

In Table 4, we choose the best LLM and BERT filtering model for each dataset based on the $F_{0.5}$ of the system. EarShot-PREDICT obtains higher precision than other systems with around 9-20 points. However, its DPR suffers obtaining around 1-13 points. While it beats MLM and EPD in terms of DPR, Word2Vec and Phrase2Vec beat PREDICT on Balanced and Synthetic and tie it on Realistic stemming from keyword extraction models, which aim to identify important words rather than specifically targeting dog whistles themselves.

To achieve this high precision, the system makes the conservative choice of predicting on average less than 1000 words and only chooses $1-$grams. While its prediction abilities are still restricted, this prediction threshold improves over the other systems as it maintains high precision.

Looking towards the choices of keyword extractors, KeyBERT does the best on smaller datasets like Balanced and Synthetic and YAKE and RAKE do better on larger datasets like Realistic. For the LLM model choices, LLaMa 13B does the best across Balanced and Realistic while LLaMa 8B does better on Synthetic. As for the BERT model choices, RoBERTa R4 does best across Balanced and Synthetic while HateXplain BERT does well

Scenario	Filtering Model	Keyword Extraction Model	Threshold	Max n-grams	Prec	DPR	$F_{0.5}$
Balanced	RoBERTA R4	KeyBERT	50	1	**14.81**	1.65	**5.70**
	LLaMa 13B	KeyBERT	100	1	11.43	1.65	5.22
Synthetic	RoBERTA R4	KeyBERT	800	1	14.86	**13.45**	**14.55**
	LLaMA 8B	KeyBERT	400	1	**19.13**	7.14	14.32
Realistic	HateXplain BERT	YAKE	3200	3	**10.00**	1.47	**4.63**
	LLaMa 13B	RAKE	200	1	9.52	1.47	4.55

Table 4: Best Results for FETCH LLM and BERT PREDICT based on $F_{0.5}$. Prec is Precision. DPR is Data Potential Recall. Best scores for each dataset are **bolded**

on Realistic. Figures detailing choices of keyword extraction models, filtering models, and thresholds can be found in Appendix F.

BERT-based models outperform LLMs by 0.1–0.5 points across all three datasets. Despite their extra parameters, larger LLMs do not surpass task-specific models in this task.

5.3.2 DIRECT

Scenario	Model	Prec	DPR	$F_{0.5}$
Balanced	LLaMa 13B	**2.97**	13.58	**3.52**
	LLaMa 8B	2.35	**21.40**	2.86
	Mistral 7B	2.87	14.81	3.42
Synthetic	LLaMa 13B	**20.31**	56.30	**23.29**
	LLaMa 8B	14.63	46.64	16.95
	Mistral 7B	18.26	**58.82**	21.18
Realistic	LLaMa 13B	0.00	0.00	0.00
	LLaMa 8B	**0.94**	60.29	**1.17**
	Mistral 7B	0.69	46.32	0.86

Table 5: Results for DIRECT Pipeline. Prec is Precision. DPR is Data Potential Recall. Best scores for each dataset are **bolded**

Table 5 summarizes the performance of the EarShot DIRECT system. Direct prompting of an LLM achieves high DPR but moderate precision, due to the model overestimating the number of dog whistles, inflating DPR at the cost of precision.

Diving into the models, LLaMa 13B does the best across Balanced and Synthetic while it fails on Realistic where LLaMa 8B does the best as LLaMa 13B overestimates the number of dog whistles.

DIRECT outperforms all except PREDICT, which surpasses DIRECT on Balanced and Realistic datasets, while DIRECT excels on Synthetic in terms of $F_{0.5}$. DIRECT, though more powerful, fails in realistic settings. However, this result has caveats: DIRECT achieves significantly better DPR, identifies many dog whistles, is less conservative than PREDICT, and is not constrained by an n-gram threshold due to its generative nature.

6 Discussion

We evaluate our best EarShot systems across datasets, highlighting their strengths and limitations. We compare differences in outputs between models and analyze challenging dog whistles to steer potential improvements beyond EarShot. We also provide analysis of novel dog whistles found by our system in Appendix G to analyze putting our method to the test in a simulated moderation scenario.

6.1 Difference in Dog Whistles

6.1.1 Synthetic

Comparing our best model to the Unigram Word2Vec model, the Word2Vec model finds very simplistic dog whistles like "pepe", "autogynophilia", "13/50", "npc". These predictions are limited to internet slang. However, our method is able to capture more complex phraseology like "middle class", "climate alarmists""entitlement program" and more often used by politicians. Comparing to the bigram Phrase2Vec, the Phrase2Vec model still chooses simple internet slang like "amerimutt" and "cuckolds". This shows that in scenarios with many dog whistles, more complex models are able to fetch more dog whistles than Word2Vec.

6.1.2 Realistic

In the Realistic setting, Earshot achieves high precision, distinguishing dog whistles from noise. It identifies only explicit terms like "illegal aliens," while Word2Vec found some (e.g., "pedophilia," "groomer") but struggles with false positives. Moreover, Word2Vec primarily retrieves variants of seed terms ("grooming", "pedes"), whereas our method uncovers deeper links (e.g., "mass migration" and "illegal aliens,"), highlighting stronger models in real-world scenarios, where Word2Vec fails.

6.1.3 Balanced

Every system has limitations. In the Balanced scenario, Word2Vec identifies subtler and novel dog whistles (e.g., "gibsmedats", "we wuz kangs", "Baltimore" "chosenites", "africanization", "vikangs", "regressives"), while our method focuses on explicit slurs ("kike", "nigger", "faggotry."). This likely results from our conservative threshold, which prioritizes performance, and a BERT filtering system that surfaces only the most overt dog whistles.

6.2 Difficult Dog Whistles

First, emoji-based dog whistles were difficult to find as no model detected symbols like the "OK" white power symbol, the milk emoji, or the SS logo as an emoji due to the challenges in tokenizing emojis, especially in models with smaller tokenizers.

Second, context dependency and deliberate ambiguity posed significant challenges. Phrases such as "Barack Hussein Obama," "Federal Reserve," "New York Elite," "Willie Horton," and "abolish birthright citizenship" were missed due to their varied contextual usage.

Third, recency presented an issue. Newer dog whistles, like "jogger" (coined in 2020 following the shooting of Ahmaud Arbery), "save women's sports;" (an anti-trans rallying cry), and "vegan cat" (an anti-trans dog whistle), were harder to detect due to evolving and shifting lexicons. Our best models usually defaulted to older and more known dog whistles. In the Realistic scenario, terms like "illegal immigrants" and "illegal aliens," were detected. The Balanced scenario identified dog whistles like "ZOG," "Baltimore," "Sharia," and "thug." In the Synthetic scenario, older dog whistles such as "George Soros," "Google," "Rothschilds," and "White Lives Matter" were predicted. However, in the Synthetic scenario, our model did identify some newer dog whistles like "genderist" and "You Will Never Be a Woman,", demonstrating promise in detecting emerging terms.

6.3 Future Directions

Our analysis highlights critical areas for enhancing dog whistle discovery. First, complex, nuanced, and recent phrases frequently escape discovery. Second, the dependence on hate speech classifiers tends to prioritize explicit slurs while overlooking subtler forms of harmful language. Third, the inherent precision-recall tradeoff presents substantial challenges, as evidenced by our system's high precision and Word2Vec's superior recall.

Additional improvements could include hybridizing Word2Vec and LLMs, post-processing noisy predictions to enhance precision, ensembling multiple LLMs, scaling model size, adding chain-of-thought reasoning (Wei et al., 2023), or integrating syntactic or semantic information.

7 Conclusion

This work lays the groundwork for detecting novel dog whistles with NLP. We introduce a new evaluation task, rigorously test state-of-the-art systems across multiple datasets, and propose a superior approach. While Word2Vec and BERT-based systems under perform, EarShot shows promise in discovering new dog whistles but needs further development for scalability and comprehensive detection. We hope FETCH! inspires future research in dog whistle discovery and advances NLP's capabilities.

8 Ethical Considerations

We identify three key ethical considerations for our work. First, dog whistles are highly culturally specific. Since our study focuses on American cultural contexts, our system may misclassify words or phrases that are not considered dog whistles in other cultures, limiting its applicability and risking false positives when applied cross-culturally. Second, our system is designed to detect hidden or coded language. This ability could have the unintended consequence of exposing language that serves to protect activism or marginalized groups. Third, there is a risk that our system may exhibit bias, disproportionately classifying language from minority communities as dog whistles. Similar systems for hate speech detection have historically encountered fairness challenges (Sap et al., 2019).

While we acknowledge these risks, we argue that deploying the system in real-world settings, with appropriate safeguards, can effectively mitigate potential harms. Specifically, we recommend incorporating human oversight during deployment to review model outputs, ensuring accountability, and minimizing the risks of false positives and biases. Furthermore, given the critical importance of addressing hate speech, we contend that the benefits of deploying a carefully managed system like ours outweigh the potential harms when implemented responsibly.

9 Limitations

We identify a couple key limitations of our paper. First, we work with LLMs that are expensive to run requiring fairly large GPUs to run experiments with. We try to limit ourselves to smaller but still high performance open source models in order to reduce the overhead of our system.

Second, while a majority of dog whistles are fairly unique and can be found using regex, the lack of a dedicated human labeled corpora hinders the accuracy of the benchmark as false positives could occur. We aimed to limit this problem by including the Synthetic scenario, where the dog whistles were found through different means outside of regex.

Third, data contamination may pose a concern, as LLMs trained on internet-scale datasets could have encountered the dog whistle dataset during pretraining, potentially inflating their apparent performance. More broadly, the models might have encountered these dog whistles in unlabeled contexts. Consequently, we cannot ensure that the models have not been exposed to these examples prior to evaluation.

Finally, our method has only been tried on English only datasets. However, dog whistles have not been studied in multilingual contexts as much as in English and there are less resources to study them.

References

Mathilda Åkerlund. 2022. Dog whistling far-right code words: the case of 'culture enricher'on the swedish web. *Information, Communication & Society*, 25(12):1808–1825.

Bethany L Albertson. 2015. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37:3–26.

Tatsuya Aoki, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2017. Distinguishing Japanese non-standard usages from standard ones. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2323–2328, Copenhagen, Denmark. Association for Computational Linguistics.

Prashanth Bhat and Ofra Klein. 2020. Covert hate speech: White nationalists and dog whistle communication on twitter. *Twitter, the public sphere, and the chaos of online deliberation*, pages 151–172.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Max Boholm, Björn Rönnerstrand, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Gregor Rettenegger, and Asad Sayeed. 2024. Can political dogwhistles be predicted by distributional methods for analysis of lexical semantic change? In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 144–157, Bangkok, Thailand. Association for Computational Linguistics.

Max Boholm and Asad Sayeed. 2023. Political dogwhistles and community divergence in semantic change. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 53–65, Singapore. Association for Computational Linguistics.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. A text feature based automatic keyword extraction method for single documents. In *European conference on information retrieval*, pages 684–691. Springer.

Chroma. 2024. chroma. https://github.com/chroma-core/chroma.

Alexandra DeLucia. 2020. Little bird. https://github.com/aadelucia/littlebird.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Kevin Drakulich, Kevin H Wozniak, John Hagan, and Devon Johnson. 2020. Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics. *Criminology*, 58(2):370–402.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Robert E Goodin and Michael Saward. 2005. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Takuro HADA, Yuichi SEI, Yasuyuki TAHARA, and Akihiko OHSUGA. 2020. Codewords detection in microblogs focusing on differences in word use between two corpora. In *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pages 103–108.

Ian Haney-López. 2014. *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *Preprint*, arXiv:2203.09509.

Robert Henderson and Elin McCready. 2018. How dogwhistles work. In *New Frontiers in Artificial Intelligence: JSAI-isAI Workshops, JURISIN, SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI, Tsukuba, Tokyo, November 13-15, 2017, Revised Selected Papers 9*, pages 231–240. Springer.

Robert Henderson and Elin McCready. 2020. Towards functional, agent-based models of dogwhistle communication. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 73–77, Gothenburg. Association for Computational Linguistics.

Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, and Asad Sayeed. 2022. Distributional properties of political dogwhistle representations in Swedish BERT. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 170–175, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jon Hurwitz and Mark Peffley. 2005. Playing the race card in the post–willie horton era: The impact of racialized code words on support for punitive crime policy. *Public Opinion Quarterly*, 69(1):99–112.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Sedrick Scott Keh. 2022. Exploring euphemism detection in few-shot and zero-shot settings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 167–172, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Justin Khoo. 2017. Code words in political discourse. *Philosophical Topics*, 45(2):33–64.

Taehoon Kim and Kevin Wurster. 2024. emoji. https://github.com/carpedm20/emoji/.

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. 2024. 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)*, pages 12493–12509, Bangkok, Thailand. Association for Computational Linguistics.

Christopher N Lasch. 2016. Sanctuary cities and dogwhistle politics. *New Eng. J. on Crim. & Civ. Confinement*, 42:159.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 184–190, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.

Abulimiti Maimaitituoheti, Yang Yong, and Fan Xiaochao. 2022. A prompt based approach for euphemism detection. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 8–12, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection. *Preprint*, arXiv:2012.10289.

Tali Mendelberg. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton University Press.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to decipher hate symbols. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Quaranto. 2022. Dog whistles, covertly coded speech, and the practices that enable them. *Synthese*, 200(4):330.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Jennifer Saul. 2018. Dogwhistles, political manipulation, and philosophy of language. *New work on speech acts*, 360:84.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2017. Automated phrase mining from massive text corpora. *Preprint*, arXiv:1702.04457.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Jherez Taylor, Melvyn Peignon, and Yi-Shin Chen. 2017. Surfacing contextual hate speech words within social media. *Preprint*, arXiv:1711.10093.

Brian P Tilley and 1 others. 2020. "i am the law and order candidate": A content analysis of donald trump's race-baiting dog whistles in the 2016 presidential campaign. *Psychology*, 11(12):1941.

José Ramón Torices. 2021. Understanding dogwhistles politics - comprender la política de los silbatos para perros. *Theoria: An International Journal for Theory, History and Foundations of Science*, 36(3):321–339.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Rachel Wetts and Robb Willer. 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.

Hao Yang, Xiulin Ma, Kun Du, Zhou Li, Haixin Duan, Xiaodong Su, Guang Liu, Zhifeng Geng, and Jianping Wu. 2017. How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 751–769.

Kan Yuan, Haoran Lu, Xiaojing Liao, and XiaoFeng Wang. 2018. Reading thieves' cant: automatically identifying and understanding dark jargons from cybercrime marketplaces. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1027–1041.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection. *Preprint*, arXiv:2402.11406.

Kangzhi Zhao, Yong Zhang, Chunxiao Xing, Weifeng Li, and Hsinchun Chen. 2016. Chinese underground market jargon analysis based on unsupervised learning. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 97–102.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. *Preprint*, arXiv:2103.16808.

A   Dog Whistles Found in Each Scenario

Scenario	List of Dog Whistles
Balanced	#genderwoowoo,#milk,#womenwontwheesht,109,1290,1488,23/16,absent fathers,affirmative action,aiden,alarmism,alarmist,all lives matter,alt-right,america first,amerimutt,amish,anointed,attacks on or imagery of jewish financial leaders,attacks on or imagery of jewish political leaders,autoandrophile,autogynephile,baby daddies,baby mama,back the blue,bankers,banksters,barack hussein obama,bear emoji,beta,big government,big pharma,bing,biological man,biological realism,black-on-black crime,blue lives matter,blueish,boogaloo,bop,broken family,burning coal,butterfly,cabal,checkered flag emoji,china virus,clown world,clownfish,coin,colorblind,controlled media,critical race theory,cuck,cultural enrichment,cut taxes,deadbeat dad,deep state,dinosaur emojis,durden,election integrity,entitlement programs,entitlement spending,every single time,fafo,family values,federal reserve,food stamps,fren,frog emoji,gangbanger,gender critical,gender ideology,george soros,ghetto,gibsmedat,globalism,globalist,goody,google,government handout,groomers,hard-working americans,he wuz a good boy,hollywood elite,identify as,illegal aliens,illegal immigrant,illuminati,inner city,international bankers,islamic extremism,islamic extremists,islamic terrorism,islamic terrorists,islamists,israel lobby,"its okay to be white",jogger,judas,jwoke,kek,khazars,law and order,lesbophobia,"lets go brandon",lgb rights,lifelong bachelor,lizard emoji,maga,male violence,milk emoji,multiculturalism,neoliberal,new world order,nibba,npc,ok sign emoji,our guy,overpopulation,oy vey,pepe the frog,personal responsibility,pine tree emoji,political correctness,power level,prefix "super",public school,pulling strings,puppet masters,quotas,race realism,radical islam,reagan,red square emoji,references to cities with large racial minority populations being overrun by crime, drugs, rodents,religious freedom,right to work,rothschilds,safeguarding,school choice,secure the border,send me,sharia law,shlomo,silent majority,single parent,skittle,skype,snowflake,sonnenrad,soy boy,special interests,spiderweb emoji,spqr,steroids,strapping young buck,take back,thug,tough on crime,trans identified female,trans identified male,tras,triple parentheses,two lightning bolt emojis,vinland,voter fraud,war on christmas,war on drugs,welfare,white lives matter,windmill,working class,xx,yahoo,ykw,zionist,zionist occupation government

| Synthetic | social justice warrior, blueish, LGB rights, working class,Islamic terrorists, centipede, cuck, TRAs, globalist,Trans Identified Male, Reagan, Islamic extremists,Pepe the Frog, single parent, illegal aliens,public school, MAGA, voter fraud, soy boy, XX,gender critical, political correctness, coincidence,cut taxes, grooming, neoliberal, big pharma, Clown World,America First, government handout, New World Order,multiculturalism, references to cities with large racial minority populations being overrun by crime, drugs, rodents,beta, Islamic extremism, quotas, family values,inner city, cabal, religious freedom, cultural Marxists,steroids, NPC, deadbeat dad, 13%, fellow white people,power level, Zionist, dual citizen, 1488, Skittle,Google, race realism, Deus Vult, puppet masters, Bing,critical race theory, secure the border,Trans Identified Female, OK sign emoji, octopus,personal responsibility, actual woman, entitlement programs,baby mama, cultural enrichment, Judeo-Christian,welfare queen, gender ideology, Rothschilds, war on terror,job creators, school choice, "states rights", anointed,majority-minorit y, shekels, Barack Hussein Obama,Remove Kebab, shoah, anchor babies, ghetto,forced diversity, judicial activism, Sharia law,Federal Reserve, middle class, Hollywood elite, welfare,deep state, George Soros, illegal immigrant,affirmative action, snowflake, thug, food stamps,radical Islam, YKW, take back, law and order,the goyim know, cultural Marxism, right to work, globalism,mass immigration, bankers, genderfree, election integrity,goy, gibsmedat, Islamic terrorism, bugman,black-on-black crime, strapping young buck, biological man,frog emoji, alt-right, Kek, womyn, silent majority,Islamists, war on drugs, autogynephile, biological realism,lifelong bachelor, groomers, based, identify as,big government, entitlement spending, our guy, Yahoo,fren, "its okay to be white", Skype,institution of marriage, every single time, absent fathers,biological woman, dindu nuffin, Blue Lives Matter, Amish,special interests, triple parentheses, 41%, Pajeet,agender, hard-working Americans, forced busing,out of wedlock, broken family, property rights,safeguarding, balance the budget, international bankers,alarmist, right to know, Illuminati, male violence,overpopulation, Jwoke, All Lives Matter, autoandrophile,pulling strings, "Lets Go Brandon", baby daddies, oy vey,cosmopolitan, freedom of association, Judas,war on Christmas, hygienic, colorblind, fatherless,identitarianism, China Virus, gender socialization,controlled media, White Lives Matter, bop, early life,poisoning the well, sex is real, gangbanger,black and orange square emojis, health freedom,Zionist Occupation Government, we wuz kangz, windmill,global elite, Kalergi Plan, dindu, preemptive strike,mixed sex, economic anxiety, car salesman, genderist,brave and stunning, jogger, prefix "super", thin blue line,Sonnenrad, banksters, troon, Trilateral Commission,vaccine safety, implicit bias, central bankers,healthy tissue, erasing women, he wuz a good boy,sex-based rights, lesbian erasure, sex not gender,lesbophobia, intact, RWDS, alarmism,rootless cosmopolitan, magapede, surgical wound,burning coal, Three Percenters, gender abolitionist,three strikes laws, superpredators, SPQR,attacks on or imagery of Jewish political leaders, Dred Scott,welfare cheats, coal burner, uterus-haver, gender fandom,string puller, Boogaloo, pine tree emoji, lizard emoji,New York elite, #GenderWooWoo, New York intellectuals,protecting women and girls, merit-based immigration policy,YWNBAW, dinosaur emojis, milk emoji, "save womens sports",biological realist, cleaning up our streets, 109, COIN,Amerimutt, tough on crime, Goody, autogynephilia, Durden,coastal elite, vegan cat, FAFO, physical removal,Khazars, Aiden, pro-family, Vinland, dual loyalty,SWPL, womanface, Willie Horton, bear emoji,attacks on or imagery of Jewish financial leaders, pilpul,A Leppo, cosmopolitan elite, Holocauster, Carolus Rex,loxism, free helicopter rides, Back the Blue,adult human female, clownfish, gender abolition,abolish birthright citizenship, bix nood, war on crime,RaHoWa, cherry emoji, two lightning bolt emojis,checkered flag emoji, autoandrophilia, AFAB trans woman,butterfly, Israel Lobby, nibba, male-friendly content,open borders for Israel, Shlomo, tiny minority of men,1290, New York values |

Realistic	#genderwoowoo,#milk,#womenwontwheesht,109,1290,1488,23/16,absent fathers,affirmative action,aiden,alarmism,alarmist,all lives matter,alt-right,america first,amerimutt,amish,anointed,attacks on or imagery of jewish financial leaders,attacks on or imagery of jewish political leaders,autoandrophile,autogynephile,baby daddies,baby mama,back the blue,bankers,banksters,barack hussein obama,bear emoji,beta,big government,big pharma,bing,biological man,biological realism,black-on-black crime,blue lives matter,blueish,boogaloo,bop,broken family,burning coal,butterfly,cabal,checkered flag emoji,china virus,clown world,clownfish,coin,colorblind,controlled media,critical race theory,cuck,cultural enrichment,cut taxes,deadbeat dad,deep state,dinosaur emojis,durden,election integrity,entitlement programs,entitlement spending,every single time,fafo,family values,federal reserve,food stamps,fren,frog emoji,gangbanger,gender critical,gender ideology,george soros,ghetto,gibsmedat,globalism,globalist,goody,google,government handout,groomers,hard-working americans,he wuz a good boy,hollywood elite,identify as,illegal aliens,illegal immigrant,illuminati,inner city,international bankers,islamic extremism,islamic extremists,islamic terrorism,islamic terrorists,islamists,israel lobby,"its okay to be white",jogger,judas,jwoke,kek,khazars,law and order,lesbophobia,"lets go brandon",lgb rights,lifelong bachelor,lizard emoji,maga,male violence,milk emoji,multiculturalism,neoliberal,new world order,nibba,npc,ok sign emoji,our guy,overpopulation,oy vey,pepe the frog,personal responsibility,pine tree emoji,political correctness,power level,prefix "super",public school,pulling strings,puppet masters,quotas,race realism,radical islam,reagan,red square emoji,references to cities with large racial minority populations being overrun by crime, drugs, rodents,religious freedom,right to work,rothschilds,safeguarding,school choice,secure the border,send me,sharia law,shlomo,silent majority,single parent,skittle,skype,snowflake,sonnenrad,soy boy,special interests,spiderweb emoji,spqr,steroids,strapping young buck,take back,thug,tough on crime,trans identified female,trans identified male,tras,triple parentheses,two lightning bolt emojis,vinland,voter fraud,war on christmas,war on drugs,welfare,white lives matter,windmill,working class,xx,yahoo,ykw,zionist,zionist occupation government

B Dog Whistle Examples

Dogwhistles are linguistically challenging and fairly niche use of words. We provide more examples from each dataset in context with explanations.

B.1 Balanced (Gab)

Soros

> "*red flag warning for texas!!! gov abbott: "all americans concerned about losing our liberty to the leftist liberal agenda need to know about* **soros** *spending big in texas.*"
> — **Anonymous Gab User**

In this quote, an anonymous user on Gab informs the platform of the spending by billionaire George Soros on liberalizing reforms. Soros is a notorious boogeyman in the right-wing spaces due to his Open Society Foundations, a very influential network of non-profits that donate to liberal and progressive causes. In addition, his Jewish heritage feeds into popular antisemitic tropes about the Jews controlling the world. This dog whistle slips through many filters due to the its ability to mask possible antisemitism under the guise of criticizing billionaire donations.

Echoing

> "*from what i read it was not primarily christian reasons jews were expelled, but for practical reasons that* **echo** *today's reasons.*"
> — **Anonymous Gab User**

In this quote, an anonymous user on Gab informs the platform of their research on the many expulsions of the Jews from countries during history. The dog whistle here is the use of echo. While echo looks extremely benign in its usage, there are more sinister reasons to choose echo over other similar words. Echo is related to the antisemitic use of the triple parenthesis ((())) around Jewish names, topics, or names of organizations thought to be owned by Jews. These triple parentheses represent that the Jewish people's actions "echo throughout history" hence the use of the term echo.

Gibs Me Dat

> "*the solution to rotten tomatoes* **gibs me dat** *100% black panthers #blaxit lawerewolf #gabfam #maga #news #auspol #speakfreely #ausfam #australia #usa*"
> — **Anonymous Gab User**

In this quote, an anonymous user on Gab insults African Americans wanting the film Black Panther to have a 100% percent approval rating on the film rating site Rotten Tomatoes. The dog whistle used here is Gibs Me Dat. Due to its extremely weird spelling and non standard usage, one might pass over it as a possible misspelling or worse dialectical usage. However, this phrase mocks African American Vernacular English, a dialect of English that African American's often use, for being unintelligent sounding despite it having a well-formed grammar and more complex rules than Standard American English.

B.2 Synthetic (Reddit)

Baltimore

> "*that's what blm apparently want, no policing of the black community. they're getting it in chicago and* **baltimore** *right now, and using their new freedom to run up the body counts to levels not seen since the drug wars of the early 90s.*"
> — **Anonymous Reddit User**

In this quote, an anonymous user on Reddit criticizes the police practices of cities like Chicago and Baltimore and the influence of the Black Lives Matter movement on these decisions. The dog whistle here is Baltimore. This dog whistle is extremely difficult to pick up because it is a regular city name. However, it is a dog whistle because it is used in a similar vein to the use of "inner cities." Politicians reference a place that often has high African American populations like inner cities and Baltimore to disparage the people living in it. On a side note, Baltimore is a wonderful city that you should visit ☺.

Coincidences

> "*funny how there's been tons of rallies before but suddenly there's all sorts of nazi imagery just so convenient for the cameras at this specific rally. probably just a **coincidence** i'm sure..*"
> — **Anonymous Reddit User**

In this quote, an anonymous user on Reddit believes that the appearance of Nazi imagery at a rally was not a coincidence and someone deliberately placed it there. The dog whistle here is coincidence. Coincidence often references the antisemitic trope that the Jews control the world and that many coincidences are actually Jewish conspiracies. This dog whistle slips under the radar as one could be referring to funny coincidences like being in the same vacation location as your friend at the same time or interesting coincidences like Mark Twain predicting his own death.

Family Values

> "*holy shit... nfl is done, **family values** households are shutting this crap off big-time. thanks for sharing!*"
> — **Anonymous Reddit User**

In this quote, an anonymous Reddit user displays their happiness that people with family values are boycotting the National Football League due to their support of LGBTQ+ rights. The dog whistle here is family values which references the fact that the user does not believe queer life styles are normal. This dog whistle is tricky to find because family values is a catch all term that could have positive connotations but in this specific context, disguises a hatred towards queer relationships.

B.3 Realistic (Twitter)

Islamist

> "***islamist** congresswoman #ilhanomar is jeopardising american foreign policy so that she can pander to her base.*"
> — **Anonymous Twitter User**

In this quote, an anonymous Twitter user displays their hatred of Congresswoman Ilhan Omar of Minnesota for her foreign policy decisions. They deride her with the term Islamist. Islamism is an academic term which encompasses all political and religious movements that believe Islam should influence politics. However, it hides the more sinister usage of calling Ilhan Omar a terrorist, an extremist, and a traitor to the United States.

Illegals

> "***illegals** coming here by the thousands every day will drain every resource there is they will over crowd the schools*"
> — **Anonymous Twitter User**

In this quote, an anonymous Twitter user displays their distain for illegal immigrants and their supposed draining of government resources. The dog whistle here is illegals. Illegal immigrants are a diverse group of people ranging from many different places around the world. However, often when people deride illegal immigrants they often talk about Latino immigrants. This meaning often slips by filters because talking about illegal immigrants is not necessarily hate speech, but in some contexts, it is used as a cudgel against Latino people as opposed to other illegal immigrants from other countries.

C   Prompts used in EarShot

EarShot Method	Prompt
LLM-PREDICT	"You are a helpful assistant participating in a content moderation task of dog whistle detection.\n" "A dog whistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. The concept is named after ultrasonic dog whistles, which are audible to dogs but not humans. Dog whistles use language that appears normal to the majority but communicates specific things to intended audiences. They are generally used to convey messages on issues likely to provoke controversy without attracting negative attention. Are dog whistles present in the following sentence?\n" "{POST}\n" "Please respond ONLY Yes or No. Do not explain your answer.\n"
DIRECT	"You are a helpful assistant participating in a content moderation task of dog whistle detection.\n" "A dog whistle is the use of coded or suggestive language in political messaging to garner support from a particular group without provoking opposition. The concept is named after ultrasonic dog whistles, which are audible to dogs but not humans. Dog whistles use language that appears normal to the majority but communicates specific things to intended audiences. They are generally used to convey messages on issues likely to provoke controversy without attracting negative attention. Does the following post contain dog whistles? If they do, please identify the dog whistles.\n" "{POST}\n" 'Please respond with a JSON for a function call with its proper arguments that best answers the given prompt. Respond in the format {"dogwhistles": list of strings}.\n'

Table 7: Prompts used in EarShot
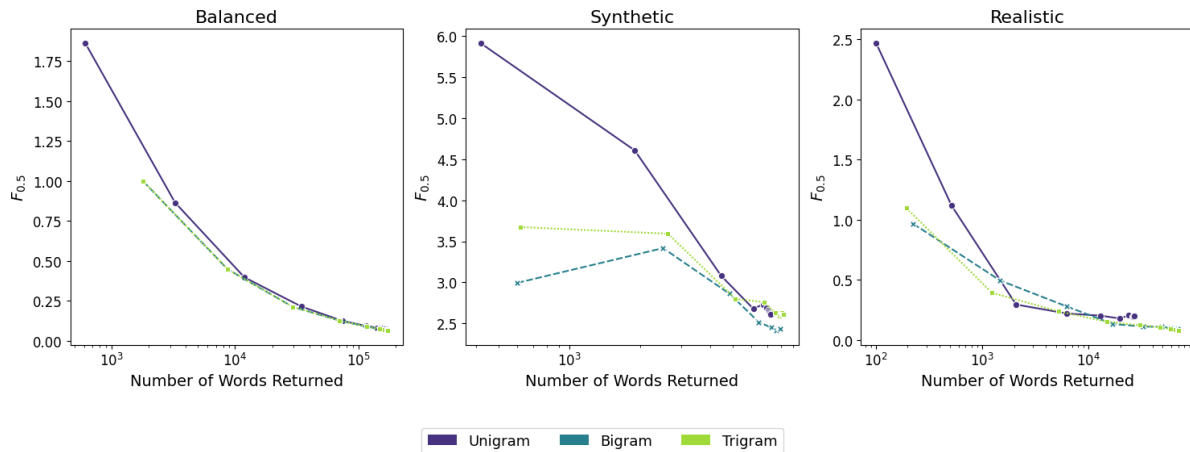
D Word2Vec/Phrase2Vec Threshold Analysis

Figure 3: Word2Vec/Phrase2Vec $F_{0.5}$ performance vs the number of words/phrases returned by the model. Plot is on log scale.
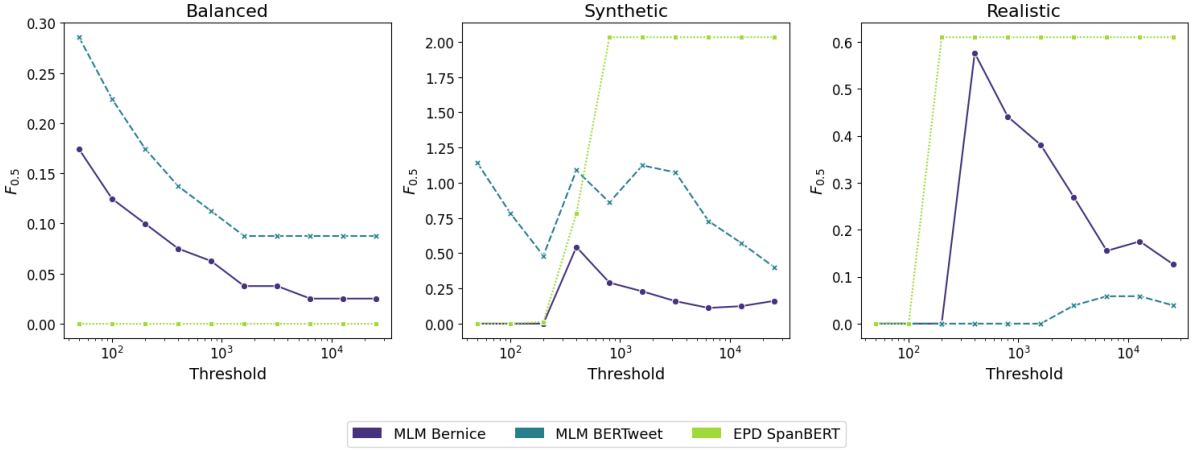
E MLM and EPD Threshold Analysis

Figure 4: MLM and EPD $F_{0.5}$ performance vs the prediction threshold. Plot is on log scale.

F EarShot PREDICT Threshold Analysis
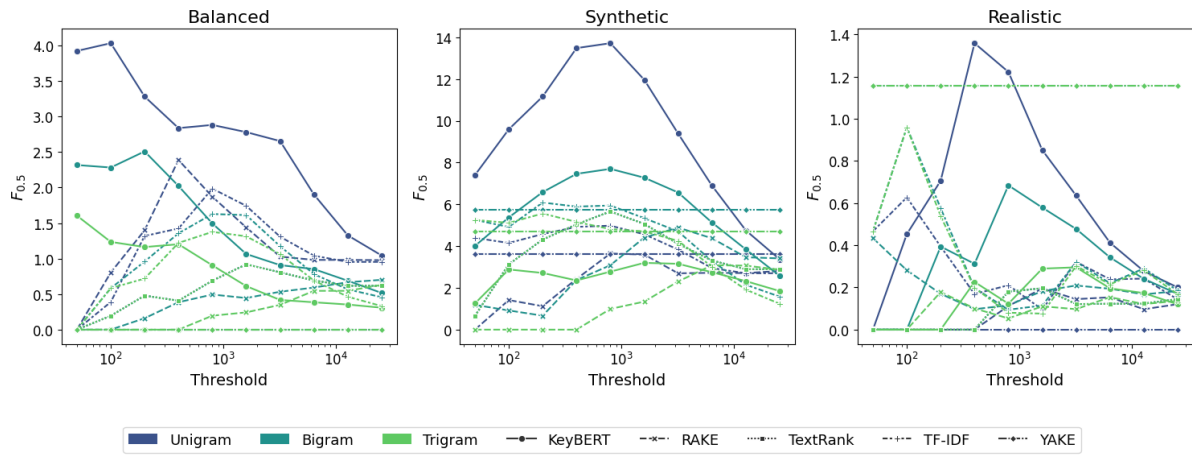
F.1 EarShot BERT PREDICT Threshold Analysis

Figure 5: $F_{0.5}$ averaged across BERT based filtering methods displaying average performance across keyword extraction models for EarShot BERT PREDICT vs prediction threshold across all three datasets. Plot is on log scale.
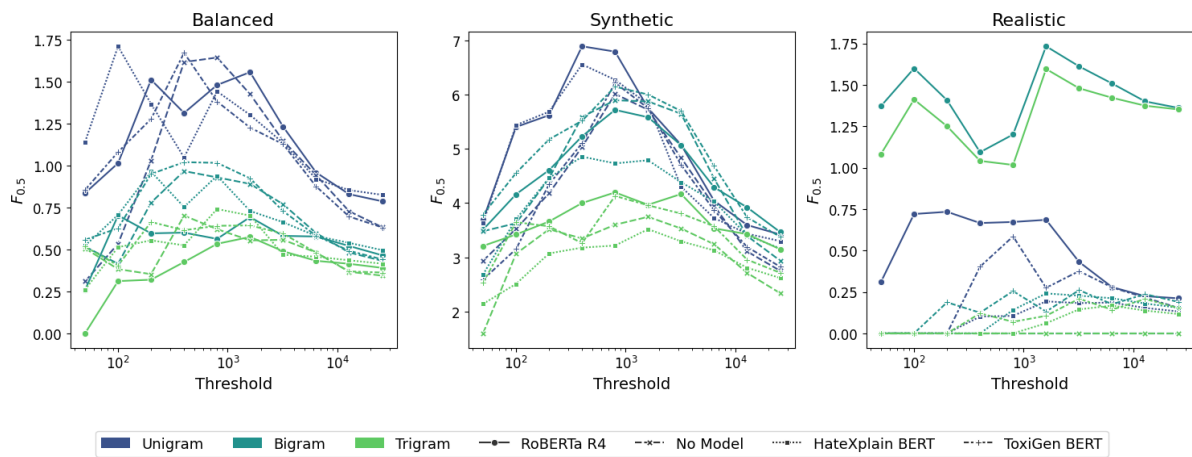

Figure 6: $F_{0.5}$ averaged across keyword extraction models displaying average performance across BERT based filtering models for EarShot BERT PREDICT vs prediction threshold across all three datasets. Plot is on log scale.

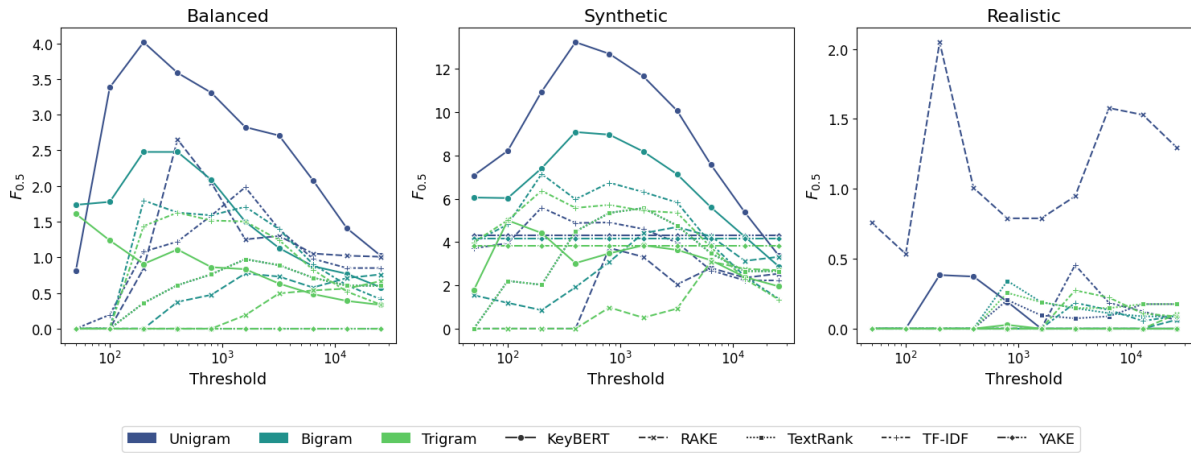## F.2 EarShot LLM PREDICT Threshold Analysis

Figure 7: $F_{0.5}$ averaged across LLM based filtering methods displaying average performance across keyword extraction models for EarShot LLM PREDICT vs prediction threshold across all three datasets. Plot is on log scale.
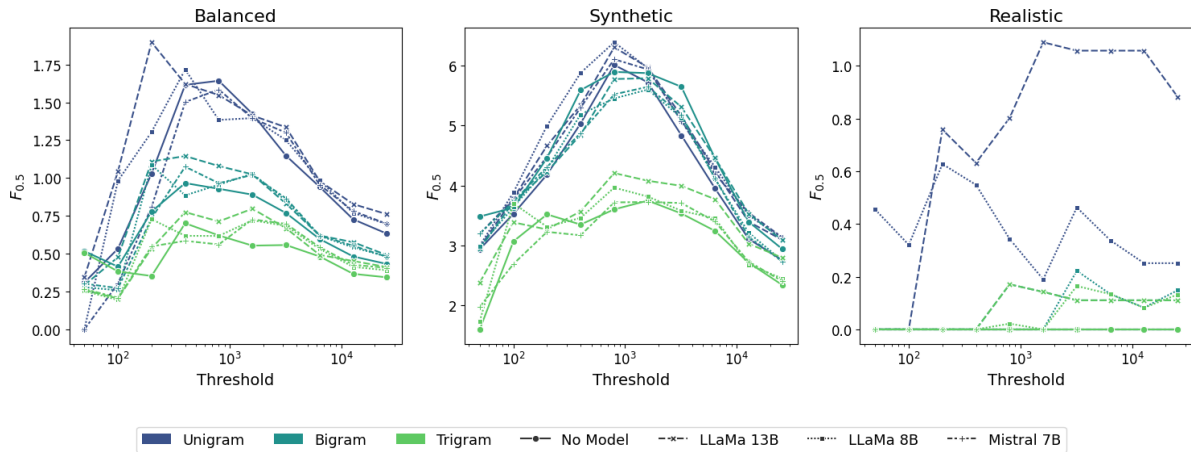

Figure 8: $F_{0.5}$ averaged across keyword extraction models displaying average performance across LLM based filtering models for EarShot LLM PREDICT vs prediction threshold across all three datasets. Plot is on log scale.

G Novel Dog Whistles

We reviewed the top systems' output to identify novel dog whistles in order to measure the effectiveness for dog whistle discovery. We detail 2–3 per dataset with more examples in Appendix G.

G.1 Balanced

The first dog whistle our model identified is "Wakanda," the fictional city from Marvel's Black Panther series. It appears in contexts suggesting African Americans should return to Africa or mocking their enjoyment of the film, as in the post: "I hate Wakandans," where it replaces the N-word. This dog whistle slips under the radar because of its benign initial contexts which could be in a discussion about the films or the comics but when used in this context it could disguise coded racism.

The second dogwhistle is "oligarchy." Used in the post: "...the #DeepState-#Oligarchy to divide and conquer us. Don't fall for their scam. We're all in this together!", it ties in with the term "deep state" which frequently implies a conspiracy about Jewish global influence. Oligarchy slips under the radar due to its multiple word senses. It could be in the post criticizing the Russian oligarchy, talking about historical oligarchies but also could hide this sense of being antisemitic.

Another novel dog whistle the model found in the Balanced scenario was brainwashing. Often, this dog whistle is used to describe either a contempt for liberal and progressive thought being supposedly being taught in public schools or Jewish control of the world. For the first meaning, the user talks about hating the supposed liberal indoctrination in schools: "true. i wouldnt let my kids go to public cesspools. and im so glad. homeschool or private school is the only answer. close the public brainwashing cesspools." For the second meaning, explicitly the user talks about the conspiracy of Jewish global control: "nope! truth isn't brainwashing, it's the antidote to filthy kike lies!" (kike being a slur for a person of Jewish descent). Brainwashing could slip under the radar as it is used in describing different policies under authoritarian regimes like Nazi Germany or describing life in religious cults.

G.2 Synthetic

One novel dog whistle found from the Synthetic scenario was "zogbots", found in the post: "...turning them into good little zogbots." This term refers to the conspiracy theory of a Zionist Occupied Government (ZOG), suggesting Jewish control making people mindless followers.

Another novel dog whistle found in the Synthetic scenario was dindu mcnuffin. It appears in the post "..impunity while blm is mad that it now shows that dindu mcnuffin did in fact point his stolen gun at the cops/innocent bystanders/puppy." This dog whistle is an modification to the previously found version "dindu nuffin" which is a contraction of "did not do nothing" attempting to sound like African American Vernacular English (AAVE). This phrase mocks AAVE and references the outcry by Black communities after a Black person is a victim of police brutality where they believe that the black person did not do nothing wrong. Dindu mcnuffin could slip under the radar as its modification of dindu nuffin which could throw off classifiers or word lists for detection. In addition, its resemblance to a name also could throw off classifiers.

G.3 Realistic

Our best system did not find any new dog whistles due to its conservative threshold.