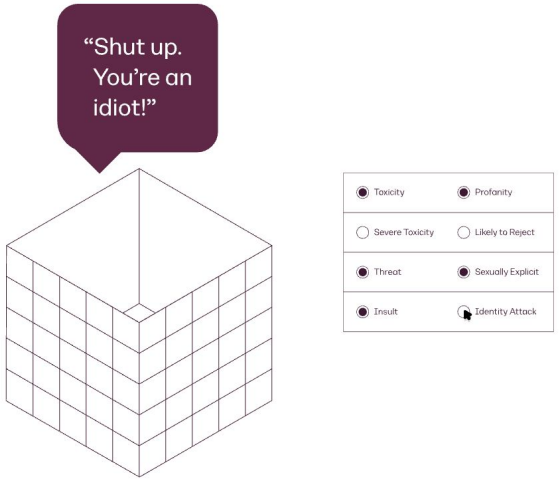
LLMs for Word-Sense Disambiguation of Coded Dog Whistles

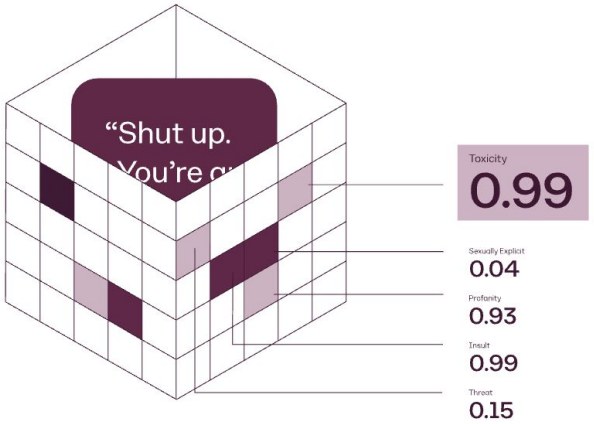
Matthew Hernandez

Motivation: prevalence of toxicity in social media

Website 1: (-) comments



Website 2: (+) comments



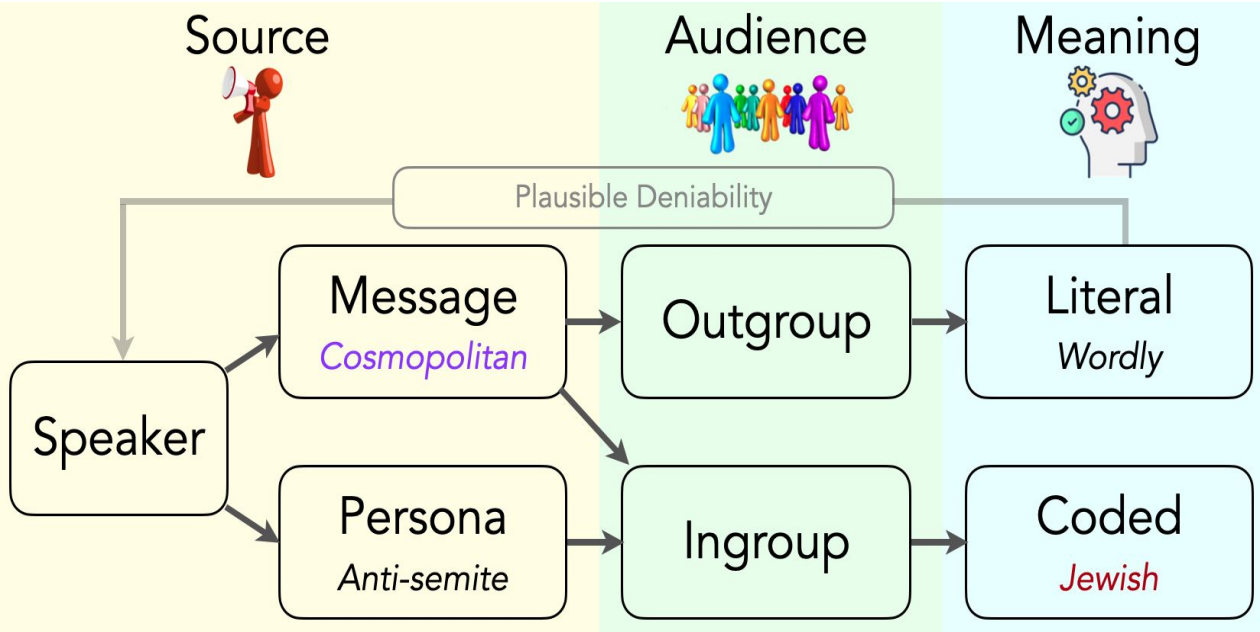
https://perspectiveapi.com/

Dog Whistles

- A form of coded language used to garner support from a particular ingroup.[1] That is, dog whistles are historically political
- They communicate harmful language but allow plausible deniability
- Difficult?
  - Yes, often undetected by NLP systems
  - They evolve over time to remain covert—exacerbated by the age of the internet
  - May look reasonable otherwise

[1] The concept is borrowed from actual dog whistles which are audible to dogs but not humans.

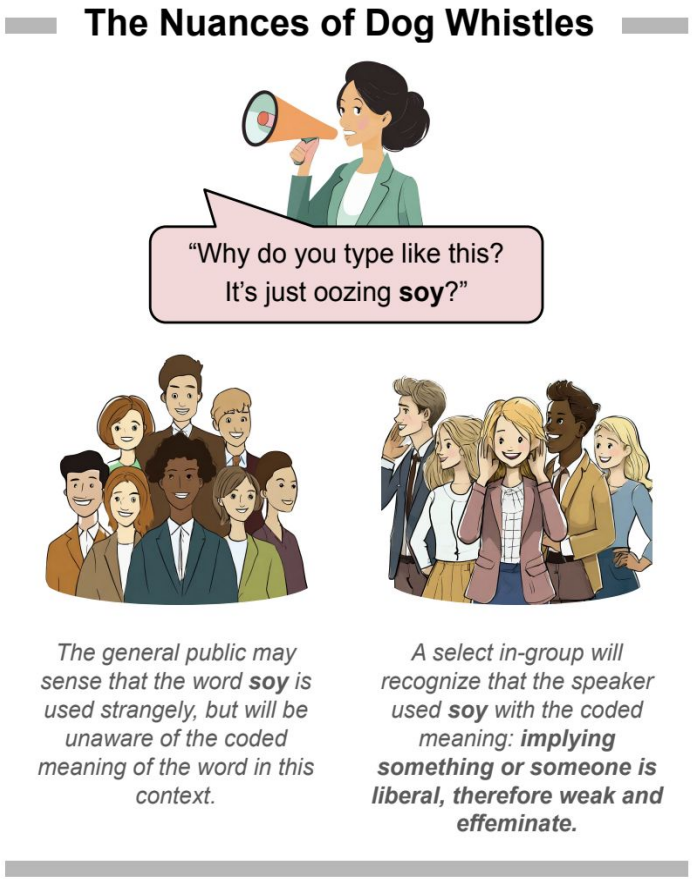
How do they work?



- Schematic based on Henderson and McCready (2018)
- [From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models](https://aclanthology.org/2023.acl-long.845) (Mendelsohn et al., ACL 2023)

Why a new dataset?

1. Study how dog whistles emerge and evolve
2. Study their prevalence in natural settings
3. Improve hate speech and toxicity detection systems
4. Unavailability of large datasets



Potential Dog Whistle Instance

- Collected from Reddit and Congressional Records
- Inventory based solely on the Allen AI Glossary of Dog Whistles
  - 340 dog whistles (lemmatized)
  - Over 1,000 surface forms
- Includes harmless instances
- Produced from [first] keyword search by using the glossary
- 327 types found

Synthetic Datasets for Evaluation

- **Synthetic-Detection**
  - 50 positive examples from Allen AI's glossary
  - 50 negative (innocuous) examples from Reddit and Congressional content[1]
  - Allen AI is the only reliable glossary of Dog Whistles[2]

- **Synthetic-Disambiguation**
  - Contains 13 unique dog whistles
  - Each type includes a set of 10 sentences
  - Structured specifically for contrastive WSD

[1] This assumes the data is balanced. Not representative of the data since dog whistles are often rare.
[2] The authors note this limitation during the evaluation of the Silent Signals dataset.

Silent Signals—Final Dataset

- Framed as a word-sense disambiguation task on the **Potential Dog Whistle Instance Dataset** by employing LLMs
- Experiments with GPT-3.5, GPT-4, Mixtral and Gemini
- These experiments demonstrated that GPT-4 can create a dataset of high confidence
- Contains 16,550 dog whistle examples.
    - Informal examples: Reddit (2008−2023)
    - Formal examples: Congressional Records (1900−2023)

Breakdown of the datasets

Dataset	Description	Size	
		Informal	Formal
Potential Instance Dataset	Produced via keyword search for dog whistle terms on data collected from Congressional records and Reddit. Used as input data for the creation of *Silent Signals*.	6,026,910	1,088,130
Synthetic-Detection	Manually annotated dataset of dog whistles examples from the *Potential Instance Dataset* used for Dog Whistle Resolution. 50/50 split on positive and negative examples.	50	50
Synthetic-Disambiguation	Manually annotated dataset where positive and negative examples are grouped by the dog whistle term they contain. Includes 13 distinct dog whistles. Designed specifically for evaluation on the Dog Whistle Disambiguation task.	74	50
Silent Signals Dataset	Novel dataset of coded dog whistle examples created by applying the Dog Whistle Disambiguation task on the Potential Instance Dataset.	13,220	3,330

Table 1: Overview of the datasets used across experiments.

Experiment 1: Automatic resolution

Prompt Design #1

- Idea is to test if LLMs are a reliable discriminator of dog whistles[1]

- **Pseudocode:**

```
FOR example ∈ examples
      LLM  ← dog whistle definition
      LLM ← candidate sentence
      RETURN binary response prediction
IF present identify the span and define it
```

Automatic Dog Whistle Resolution

You are a helpful assistant participating in a content moderation task of dog whistle detection.

A dog whistle is { **definition of *dog whistles*** }.
Does the following sentence contain a dog whistle?
If it does, please identify the dog whistle.

{ **sentence** }

Please structure your response as a JSON object, where { **structure instructions** }.

[1] Should an LLM be reliable, annotation can be done between experts and LLMs.

Results

		Human	Zero-shot				Few-shot			
			GPT-3.5	GPT-4	Mixtral	Gemini	GPT-3.5	GPT-4	Mixtral	Gemini
Presence *"is a dog whistle present?"*	Acc	66.8	80.0	**85.0**	68.0	81.0	76.0	86.6	81.0	**86.7**
	F1	64.8	83.1	**85.7**	61.9	80.0	76.0	87.4	80.0	**88.3**
Identification *"identify the dog whistle."*	Acc	49.0	58.0	59.8	59.0	**69.7**	65.7	71.1	69.0	**75.5**
	F1	33.6	56.3	48.0	45.3	**61.5**	61.4	68.2	62.7	**76.0**
Definition *"define the dog whistle"*	Acc	47.3	52.0	54.6	58.0	**66.7**	60.6	67.0	67.0	**73.5**
	F1	29.7	46.7	37.1	43.2	**56.0**	53.0	61.9	59.3	**73.5**

Table 2: Metric scores on the *Automatic Dog Whistle Detection* task which surveys LLM and human ability to detect and define dog whistles in context. When presented with a sentence these experiments test the ability of a model/user to determine if the sentence contains a dog whistle and if so, correctly identify and define it. Predictions across all models have a statistical significance of $p < 0.01$ by chi-squared test, and human predictions have statistical significant of $p <= 0.037$.

Results (cont’d)

- Human baseline (720 ex)
- For many instances the model correctly predicted a dog whistle’s presence but incorrectly identified provocative, but **non-coded**, language
- Similarly the model may correctly predict the presence of a dog whistle and correctly identify it but be unable to define it
    - Dog whistles are not ordinary definitions

Experiment 2: Word-Sense Disambiguation

Prompt Design #2

- Evaluate the LLMs' capability to distinguish contexts containing harmful coded use and ordinary use

- **Pseudocode:**

```
For dog whistle ∈ Dog Whistles
      LLM  ← wiki definition
      LLM ← 10 candidate sentences
      RETURN classification output
IF present label the span and explain why
```

Dog Whistle Disambiguation

You are a helpful assistant participating in a content moderation task of dog whistle detection.

A dog whistle is { **definition of *dog whistles*** }. The coded meaning of { **dog whistle D** } is: { **definition of D** }. Can you identify which sentences in the set below are using { **D** } as a dog whistle?

[{ **sentence 1** },
{ **sentence 2** },
…
{ **sentence 10** }]

Please structure your response as a JSON object, where { **structure instructions** }.

[1] Should an LLM be reliable, allocation can be done between experts and LLMs.

Results

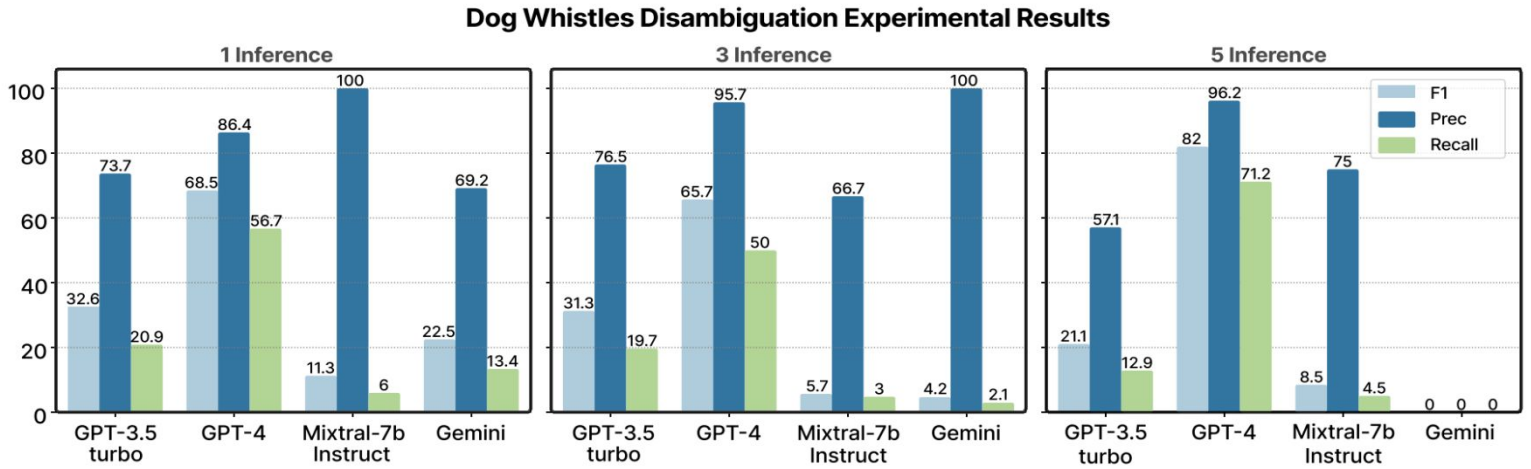


Figure 3: Results of *Dog Whistle Disambiguation* task using the simulated ensemble across $N = 1, 3, 5$ inferences. In an attempt to compensate for output volatility, for each N-inferences experiment, predictions are only considered if they remained consistent across all N runs. Precision-1 and Recall-1 scores pertain to the positive class of coded dog whistle instances.

Results (cont'd)

- Gemini and Mixtral were reluctant to generate output regarding offensive content
- Gemini's performance drastically decreased after more inferences
- The task is optimized for precision over recall[1]
  - Suggests GPT-4 is reliable enough to generate a dataset with high confidence given the WSD framework
- As a moderator, what would you optimize for?

[1] It appears the evaluation on extrinsic task was chosen beforehand, making it a good example of generating data.

Silent Signals Dataset

Silent Signals

- Addresses the limitations on a lack of trainable data for dog whistles
- Leverages the WSD methodology over 100,00 instances over the **Potential Instance dataset**
- Generated by an ensemble approach over 3 inferences with GPT-4
- Each example annotated with their respective characteristics

Validation

- Manually evaluated a sample of 400 instances using prompt #2. What is less clear is how many dog whistles were used
- The vetting procedure found a **precision** of 85.3%

- However, a number of False Positives were correct but the coded meaning was not in the Allen AI Glossary
- Considering these novel examples the **accuracy**[1] increased to 89.4%

[1] The authors wrote accuracy however I believe they meant precision.

Limitations

- No obvious baseline for WSD of dog whistles
- Would the **most frequent baseline** be useful here? Maybe source from urban dictionary? TF-IDF is applicable here too.
- Multifaceted problem
  - Size of ingroup
  - Backlash if found out
  - Are all all coded instances bad?

- However, the paper highlights potential extrinsic tasks (e.g., hate speech detection, neology, and political science). Awesome! 👍

Thank you!

Analysis

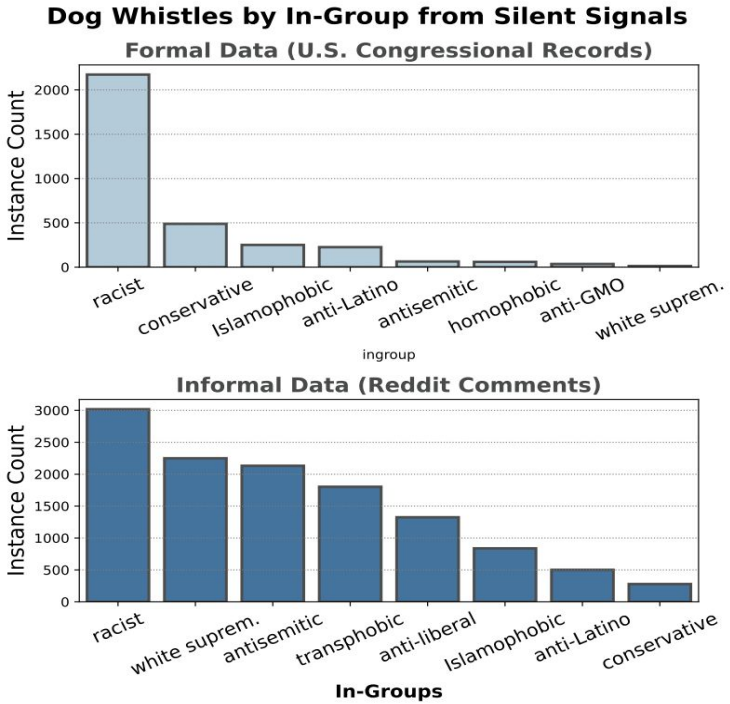


Figure 4: The distributions of dog whistles over in-groups for informal and formal communication in the Silent Signals dataset.

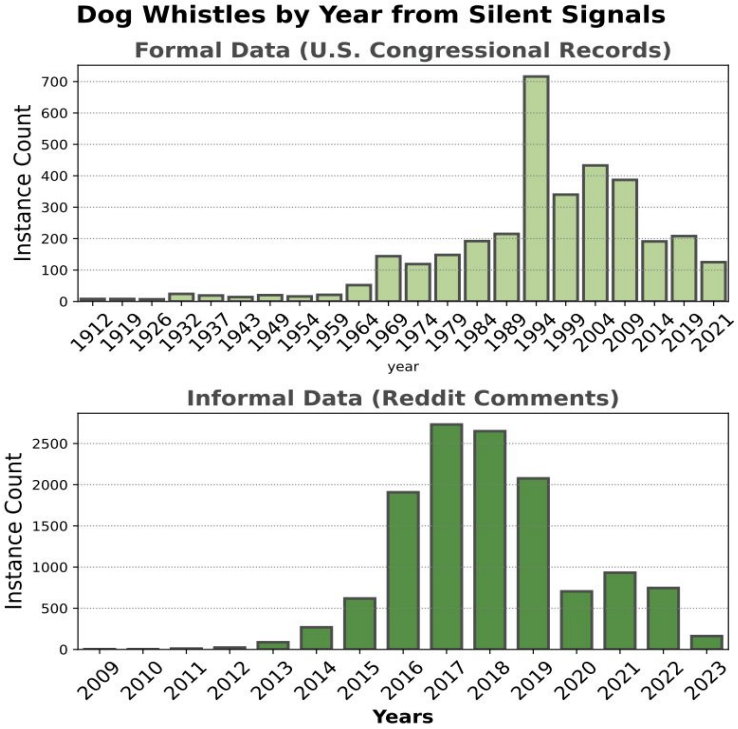


Figure 5: The distributions of dog whistles over time for informal and formal communication in the Silent Signals dataset.

Analysis (cont'd)

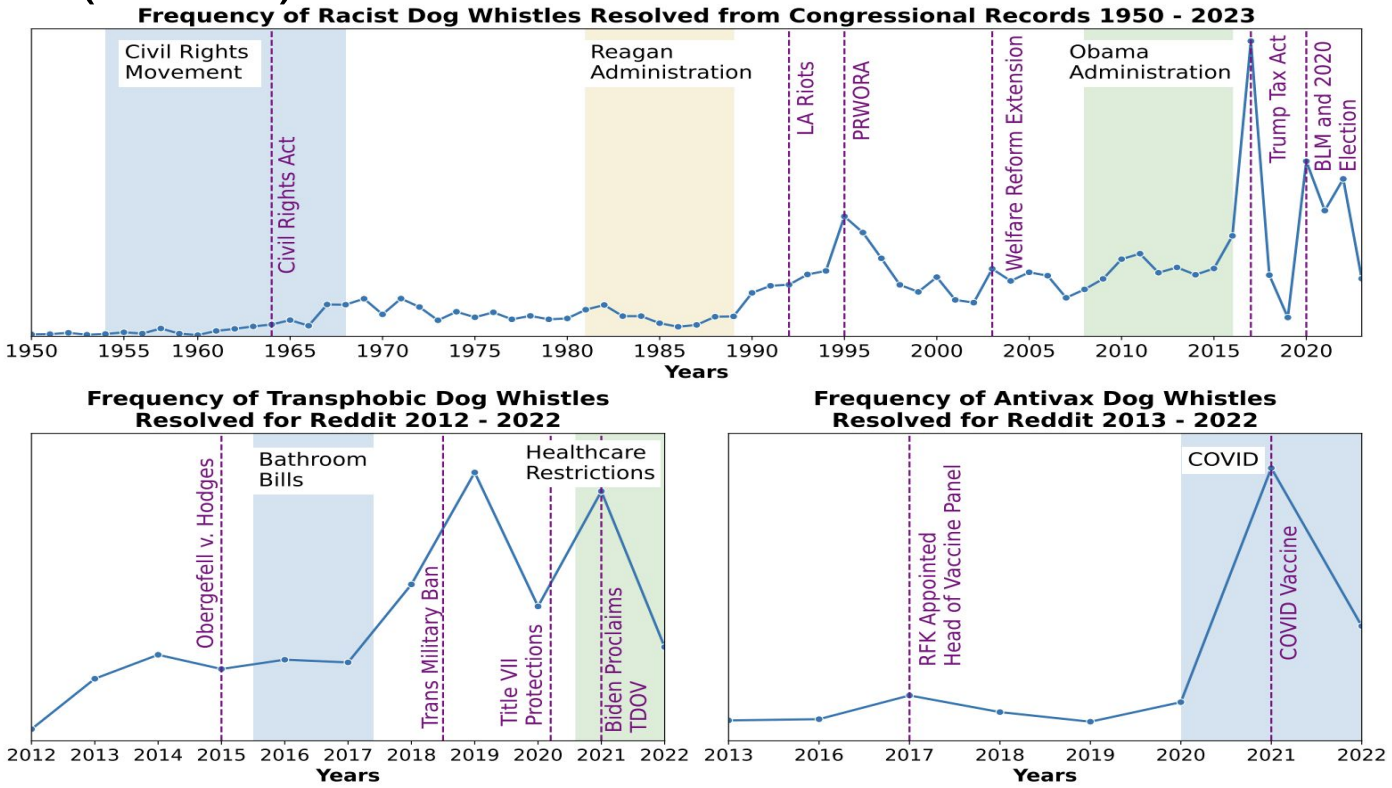


Figure 6: We investigate the use of Racist, Transphobic, and Anti-Vax dog whistles captured by the Silent Signals dataset over time. The graphs in this figure aree annotated with dates of pivotal political and cultural events in the United States.