Sam Cohen
PPOL 6821
Professor Ziogas
November 10, 2024

Final Project – Stage #5

Research Question: Can LLMs detect dog whistles accurately in social media posts using everyday speech? Will they be able to differentiate between posts using everyday terms and posts containing terms used explicitly for dog whistling?

Background

As technological methods of communication evolve, so too does the language we use to communicate through them. The invention of the radio, phone, and world wide web have all led to new words, phrases, and idioms we take for granted in everyday parlance. For instance, "hello" was not a widely used greeting until the invention and proliferation of the telephone at the turn of the century (Grimes 1992), and in our modern world, the texting-related initialism "LOL" is almost as ubiquitous. However, mediums like these also have the potential to generate new harmful or malicious language. With the rise of social media, users intending to signal bigoted or hateful opinions to other likeminded individuals may cloak their true intentions in speech that may seem harmless at first glance— a practice known as "dog whistling."

Dog whistles are not new, but with the broad reach of social media platforms like X (formerly Twitter), a tumultuous political environment in the United States, and the rise of far-right movements across Europe and North America, their prevalence is becoming increasingly visible. For this project, I will attempt to train a model to detect dog whistles. By doing so, I will: (1) try to create an algorithm to uncover language with prejudicial motivation that may not be obvious to the unassuming reader, and; (2) see if this model can tell the difference between posts that may *contain* words found in dog whistles, and those that *intend* to be harmful.

Literature Review

One of the most important studies on using neural networks and large language models to detect dog whistles was conducted by Mendelsohn, et. al. Perhaps their most consequential discovery was that LLMs like ChatGPT often fail to detect the malicious intentions of dog whistles (Mendelsohn et. al., 2023). Wetts and Willer find that exposure to dog whistles may also even influence policy views of certain individuals (mainly white liberals, in their study) compared to baseline (Wetts and Willer, 2019), and Drakulich et. al. find that dog whistles may even have impacted the choice to vote in the 2016 US Presidential election by appealing to racial biases (Drakulich et. al., 2020). One of the most important developments in this domain is Kruk et. al.'s "Silent Signals" data set, which contains thousands of disaggregated dog whistles and is currently used in hate speech detection software in various different fields (Kruk et. al., 2024).

Data: EDA and Preprocessing

I utilized Kruk et. al.'s Silent Signals data set in this model. This set contains "16,550 high confidence coded examples of dog whistles," and is available on HuggingFace, a repository of ML and AI models (Ibid). These dog whistles are taken from the website Reddit, and there are 47 unique subreddits represented in this set. The metadata contains information on the dog whistles themselves, the posts or content they were used in, the date, their etymology or origin, the "in group" associated with them, the individual or entity they intend to target, and the platforms they are taken from. Each observation in this set is taken from an instance where the dog whistle in question was used. There are over 700 hundred unique dog whistles in this set, and over 16,000 unique instances of posted content in which they are used.

After some exploratory data analysis, it was discovered that some of the most common dog whistles related to white supremacist groups (2,229). This was followed closely by dog whistles that were antisemitic (2,156), transphobic (1,807), "anti-liberal" (1,255) and Islamophobic (1,075). The subreddit in the set with the highest usage of dog whistles as "r/The_Donald," the page dedicated to former-President Donald Trump.

While no reddit post features contained missing values, several other features did, namely: speaker, chamber, subreddit, and political party. Due to these variables being of secondary importance to their associated posts, imputation or row deletion may not be necessary. However, reformatting, stripping and text analysis was completed for the posts in question.

One significant caveat was that the Silent Signals data set did not contain any observations for the negative class. That is, it only contained posts with dog whistles, meaning that supplemental data was necessary to generate the noise needed for a predictive model. To account for this, reddit post text data from other subreddits was used, namely from Question and Answer (Q and A) pages. A data set found on HuggingFace posted by username "nreimers" titled "reddit questions best answers" was used, containing reddit posts on questions being asked, the body of the question (a longer form explanation of the question being posed), a score (indicating interactions like upvotes and downvotes), and answers to the question. There were over 1.83 million individual observations in this repository, and the dates of each post ranged from 2010 to 2013. Ultimately, a subsample of 16,000 posts from this repository was used, roughly matching the number of dog whistle posts.

My reasons for choosing this particular data set were twofold. For one, users are not likely to use dog whistles purposefully in these posts-- in addition to the fact that the subject matter is usually more "tame," they are merely asking and/or answering questions in good faith. That is, even if certain words akin to dog whistles are used, they are likely not intended to be dog whistles. Secondly, Q and A forums are some of the most popular and widely used pages on reddit (Reddit 2024), so inclusion may give the model a significant amount of common subject matter and language from hundreds if not thousands of focus areas.

Research Design

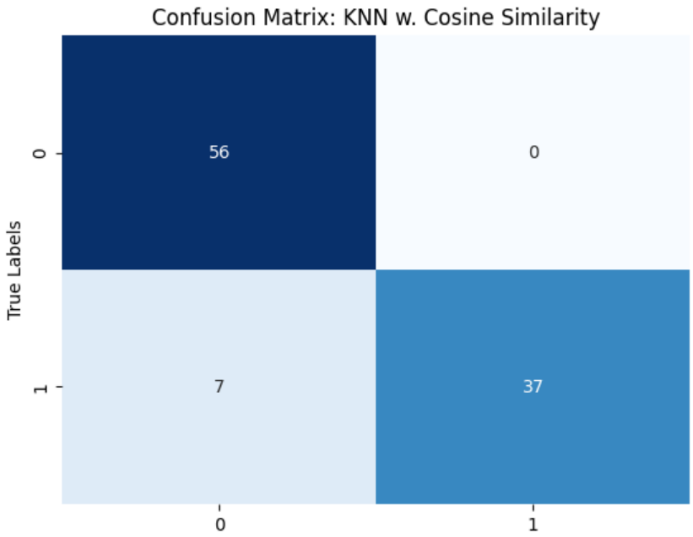
To leverage the power of LLMs, I used a pre-trained language model to generate word embeddings for the training data (i.e. both posts containing dog whistles and posts not containing dog whistles). The target variable in this case was whether or not a post did (1) or did not (0) contain a dog whistle. A test set threshold of 20 percent of the observations was set aside for validation.

First, word embeddings were generated from a pretrained model and mapped to each token for each post in the training data. The language model employed was DistilBERT Uncased, a smaller but more compact and faster iteration of BERT (Bidirectional Encoder Representations from Transformers), a detailed, pre-trained LLM that takes token placement into account (Hugging Face 2024). This model was used due to its relatively quick training time, simplicity, and efficiency.

Once embeddings were generated, I used a K-Nearest Neighbor model (KNN) with a cosine similarity distance metric to classify test posts. This involved checking the distance of the token embeddings in a test post to the embeddings of those in the training data. If the majority of embeddings in the test post were closer to K=3 amount of dog whistle posts than they are to non-dog whistle posts, it will be classified as likely containing a dog whistle. I then compared these results to that of a simple logistic regression to see which garnered a higher test accuracy score.

Results

After fitting the KNN model to the training data, a five-fold cross validation mean accuracy score of 82.73 percent was obtained. A similar score of 82 percent was obtained after feeding test data through the trained model. However, after changing the distance metric to cosine similarity (a common practice for word embeddings and text classification models), the five-fold cross validation mean accuracy score climbed to 94.6 percent, with a test accuracy score of 93 percent. It should be noted that to account for the time it took to generate new embeddings, only the first 100 posts of the test data went through the word embedding process. The model predictions for the test data were then compared to the first 100 true labels of the test target data. A confusion matrix indicated a true negative rate of 56 percent and true positive rate of 37 percent. This model was therefore better at classifying non-dog whistle posts. The false negative rate of 7 percent, meaning that 7 percent of test posts were misclassified as dog whistle free. The false positive rate was negligible.



Above: Confusion matrix for KNN on test data (generated embeddings)

To see how other algorithms compared to the KNN model above, I also fit the training data to a logistic regression. This model resulted in an accuracy score of 96 percent, and a much lower false negative rate of 2 percent. The false positive rate increased slightly to 7 percent, though for the purposes of this model, a higher false positive rate is preferable to a high false negative rate, since posts incorrectly classified as "safe" may lead to their inability to be flagged.

Finally, to compare the generated word embeddings to other pre-existing ones, I generated sentence embeddings using Word2Vec, a shallow network word vectorizer (Pathmind 2024). Word2Vec can generate word embeddings at a much less computationally expensive rate, but cannot take into account nuance or context as well as transformers, making it an ideal foil for the transformers. After getting sentence embeddings for each post in the training data, a logistic regression model was fitted. After feeding in the test data, an accuracy score of 92.8 percent was the result. Interestingly, KNN performed the least well with worc2vec, with an accuracy score of 84 percent.

Evaluation

a. Limitations

Perhaps the greatest initial limitation I encountered was how to deal with missing target class data. Because the dog whistle data set from Hugging Face only contained the positive target class (i.e. posts with dog whistles), additional data were needed to provide noise. While the choice of Q and A post data ultimately did provide noise, the optimal data set would have included many posts that have similar language and structure to the dog whistle posts, without explicitly including dog whistles or malicious intent themselves. This would help the models learn even more nuance. However, finding enough data to fit this criterion would be exceedingly tedious and time consuming.

Overfitting is also a possibility here. The primary reason original word embeddings were generated was to provide a model that was specifically tailored to dog whistles. However, in doing so, the model may be too adept at its original task and intention. That is, if posts are structured slightly

differently, or new dog whistles come about, the model might have a hard time detecting them. On the contrary, the model might also have the potential to classify posts that have similar language to dog whistles but do not actually contain any. For instance, if a reddit post contains hateful rhetoric but no actual dog whistle, the model may classify it as containing one anyway.

Another barrier was long training times and computational expense. When creating word embeddings for the training posts using DistilBERT, the process took hours, even when leveraging the Colab GPU. A Google Colab Pro subscription was needed to resolve this.

Finally, there are temporal limitations to this model. Because the data used ranged in dates from 2016 to 2023, future iterations of this model should focus on dog whistles that may come about over the next several years.

b. Reflections

Overall, the models created performed well on the data they were given. However, models are only as useful as the quality of data they are fed, so results must be taken with a grain of salt, even though each model was able to classify several posts from outside sources correctly,

These models presented a unique challenge for me in the field of text classification. They employed a combination of traditional supervised machine learning methods (i.e. KNN and logistic regression) as well as neural networks and LLMs. While other algorithms like naïve bayes may have provided a relatively useful and less computationally expensive alternative, in the specific case of dog whistles, it is not only the tokens but language and nuance surrounding them that are key to classification.

Though the models developed here may not be perfect, they are a step in the right direction, especially as social media platforms become less regulated. Future research and implementation should focus on taking into account new dog whistles that develop overtime in the online world, as well as different mediums, such as different neural networks that can classifying dog whistles in audio and speech patterns.

Bibliography

DistillBERT (2024). DistilBERT Uncased. Hugging Face. Accessed October 2024.

https://huggingface.co/distilbert/distilbert-base-uncased

Drakulich, Kevin et. al. (February 2020). Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics. Criminology, Vol. 58(2). https://doi.org/10.1111/1745-9125.12239. https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9125.12239

Grimes, William (1992). Great "Hello" Myster is Solved. New York Times.

https://www.nytimes.com/1992/03/05/garden/great-hello-mystery-is-solved.html

Glossary of Dog Whistles (2023). Association for Computational Linguistics.

https://dogwhistles.allen.ai/glossary

Kruk, Julia et. al. (August 2024). Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles. Association for Computational Linguistics, Vol. 1. https://aclanthology.org/2024.acl-long.675.pdf

Mendelsohn, Julia, et. Al. (July 2023). From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. Association for Computational Linguistics, Vol. 1. https://aclanthology.org/2023.acl-long.845.pdf

Meta Research. Introducing LLaMA: A foundational, 65-billion-parameter large language model. February 24, 2023. Accessed October 14, 2024. https://ai.meta.com/blog/large-language-model-llama-meta-ai/.

Nicholson. Chris. (2023). A Beginner's Guide to Word2Vec. Pathmind. Accessed November 2024.

A Beginner's Guide to Word2Vec and Neural Word Embeddings | Pathmind

Nreimers (2024). Reddit Questions Best Answers. Hugging Face. Accessed October 2024.

nreimers/reddit_question_best_answers · Datasets at Hugging Face

Reddit (2024). R/communities. Accessed November 2024.

https://www.reddit.com/best/communities/1/

SALT-NLP (Accessed October 2024). Silent Signals. Hugging Face.

https://huggingface.co/datasets/SALT-NLP/silent_signals

Sayeed, Asad et. al. (February 2024). The utility of (political) dogwhistles – a life cycle perspective.

University of Gothenburg. https://doi.org/10.1075/jlp.23047.say. The utility of (political) dogwhistles – a life cycle perspective | John Benjamins (jbe-platform.com).

Silent Signals Dataset. SALT-NLP/silent_signals. HuggingFace. Accessed September 16, 2024.

https://huggingface.co/datasets/SALT-NLP/silent_signals

Wetts, Rachel and Willer, Robb (August 7, 2019). Who Is Called by the Dog Whistle? Experimental

Evidence That Racial Resentment and Political Ideology Condition Responses to Racially Encoded Messages. American Sociological Association, Vol. 5. https://doi.org/10.1177/2378023119866268. https://journals.sagepub.com/doi/10.1177/2378023119866268?icid=int.sj-abstract.citing-articles.15