Distributional properties of political dogwhistle representations in Swedish BERT

Niclas Hertzberg* and **Asad Sayeed*** and **Ellen Breitholtz*** and **Robin Cooper***
and **Elina Lindgren†** and **Gregor Rettenegger†** and **Björn Rönnerstrand†‡**
*Dept. of Philosophy, Linguistics, and Theory of Science
†Dept. of Journalism, Media, and Communication
‡SOM Institute
University of Gothenburg, Sweden
asad.sayeed@gu.se

Abstract

"Dogwhistles" are expressions intended by the speaker to have two messages: a socially-unacceptable "in-group" message understood by a subset of listeners and a benign message intended for the out-group. We take the result of a word-replacement survey of the Swedish population intended to reveal how dogwhistles are understood, and we show that the difficulty of annotating dogwhistles is reflected in the separability of the space of a sentence-transformer Swedish BERT trained on general data.[1]

1 Introduction

We explore whether contemporary vector-space sentence representation techniques also provide a structured representation of the different messages in "dogwhistle" political communication. A dogwhistle refers to a word or phrase used in manipulative communication, usually in a political context. Dogwhistles carry at least two messages: one message intended for the broader community, and another "payload" message intended to communicate a specific, less acceptable message to a receptive "in-group". Dogwhistles depend on the "out-group" members not picking up on the payload message (Albertson, 2014; Bhat and Klein, 2020).

We take several Swedish-language dogwhistles and survey data from the Swedish population about the interpretation of these dogwhistles, and we apply clustering techniques based on the transformer-derived representation of the responses. We ask the question: are the responses clearly partitioned in the semantic space, and does the "sharpness" of this partitioning reflect the ease of dogwhistle identification by expert annotators?

While there has been work exploring dogwhistles through the lens of linguistics (Henderson and McCready, 2019; Bhat and Klein, 2020; Saul, 2018), automated approaches to exploring dogwhistles using NLP techniques are generally lacking (Xu et al., 2021). Considering the volume of social media data and the extent to which dogwhistles have been employed on these channels, it is important to create new computational techniques to detect and analyze dogwhistles that might succeed at higher data volumes. The first step in accomplishing this is to show that automatic techniques can be used to reliably extend and enhance manual analysis.

Dogwhistles can be strategically used, e.g. politically to send a veiled message to one group of voters while avoiding alienating another group (Bhat and Klein, 2020). This could pose a problem in a representative democracy since the out-group portion of the voter-base are deceived into voting for a certain candidate that might not represent their political views (Goodin and Saward, 2005).

Therefore, we contribute the following:

- We present a preliminary dataset of a word replacement task by members of the Swedish population as part of a survey of political attitudes, including a manual annotation for dogwhistle identification with inter-annotator agreement (IAA; Krippendorff's α) scores.
- We use a transformer-based model to represent the responses in a semantic space and apply classification (SVM) and clustering techniques (K-means) to the vectors.
- We evaluate the clusterings in terms of cluster purity metrics, and we show that the lower the IAA, the lower the linear separability of the responses in the vector space.

We then conclude that a Swedish BERT variant already represents important aspects of the underlying semantics of dogwhistles.

[1] Authors other than Niclas Hertzberg and Asad Sayeed are listed in ascending alphabetical order.

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 170 - 175

2 Dataset

Dogwhistle politics has become increasingly salient in the current mass and social media environment. This is also the case in Swedish society. Recent studies have shown that certain issues, in particular immigration, have produced examples of emergent dogwhistles gaining in public use (Åkerlund, 2021; Filimon et al., 2020).

Using a professional polling firm, we anonymously sampled 1000 members of the Swedish public using a word replacement task. We constructed 5 sentences containing words or phrases we suspected were being used as dogwhistles and asked survey participants to replace the words with what they thought it "really" meant. Then we manually annotated these responses for whether they identified a dogwhistle use or not. The survey was conducted under institutional ethical review in a process that involved survey administration and anonymized data compilation at a remove from the authors.

Each item therefore contains the substitution of participant-provided words or phrases for the original dogwhistle in the full context of the corresponding stimulus sentence. An illustrative stimulus example would be the following: "The Swedish unions are controlled by **globalists**". Each person taking the survey would replace “globalists” with a word or phrase they believe to convey the same information. The replacements can vary widely: someone might replace “globalists” with "communists" or an anti-Semitic slur, which might be considered an "in-group" response. Others would replace “globalists” with, e.g., "people concerned with international affairs" thus not showing an understanding of the dogwhistle as having any associations with the aforementioned groups. The actual Swedish dogwhistles we use and their English translations are listed in table 1.

Each replacement thus gave rise to a slightly altered sentence that, according to the person taking the survey, would convey the same information as the original sentence. The replacements for each dogwhistle was manually labeled depending on a person picking up on the dogwhistle meaning or not. An inter-annotator score was then calculated for the labeling of each dogwhistle.

IAA was calculated in two rounds, an initial round and a confirmatory round partway through the annotation. We report both scores in table 2.

Role of inter-annotator agreement[2] The goal of the annotation and the computation of IAA is to determine whether or not the annotation task can be designed with the following criterion in mind: that a panel of trained annotators with access to the guidelines can reliably distinguish between participant responses that *did* pick up on the "in-group" dogwhistle meaning from those that *did not*.

The identification and interpretation of a dogwhistle is an inherently subjective task which stems directly from one of the reasons to use a dogwhistle in the first place: to take advantage of the ambiguity of interpretation based on the standpoint of the individual recipients of the message. There are good reasons to critique the widespread use of IAA statistics to represent reader or listener reaction in subjective tasks like these (Sayeed, 2013). However, in this case, the annotation guidelines were developed in an iterative process to be presented in future publications that ensured that Swedish-speaking annotators informed about Swedish politics could consistently identify the dogwhistle interpretations of survey participants. The focus of this work is to explore the extent to which the intuitions behind the annotation guidelines are reflected in a Swedish BERT model trained on a multi-genre corpus.

3 Method

3.1 Sentence transformers

Sentence transformers (Reimers and Gurevych, 2019) are based on BERT (Devlin et al., 2018) and produce state of the art semantic representations of entire sentences and paragraphs. A high performing sentence model returns semantic representations of sentences, with a cosine distance that correlates with their semantic similarity. Different sentences can thus be compared computationally. The specific sentence model we used was Swedish sentence-Bert (Rekathati, 2021).

Resources for training machine learning models on Swedish text are somewhat limited. The lack of resources prevents training a sentence transformer in Swedish using the same procedure as training sentence transformers in English. However, the training of a sentence transformer in the target language can be obtained by fine-turning a Swedish model (Malmsten et al., 2020)[3] on the output of an

[2] We thank Reviewer 3 for raising this point.

[3] Pre-trained on books, newspapers official government reports, a small amount of social media, and Swedish Wikipedia.

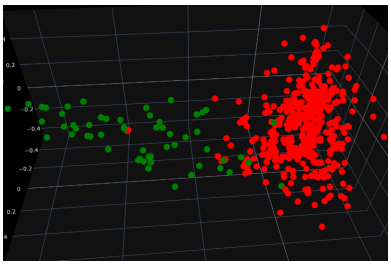

Figure 1: Responses for dogwhistle "enrich" represented in the semantic space. Coded in-group responses colored green.

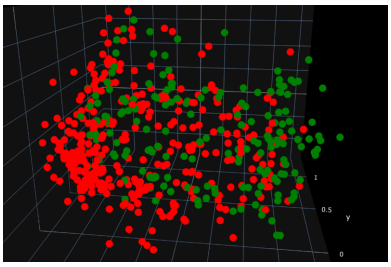

Figure 2: Dogwhistle "remigration" represented in the semantic space. Coded in-group responses colored green. Additional plots are in supplementary material.

already trained English sentence transformer and a parallel corpora of the source and target language. (Reimers and Gurevych, 2020). This procedure is an accessible way to train sentence transformers in a variety of languages faced with the same data limitations as Swedish.

3.2 Procedure

As we were interested in the semantic representations given by the sentence replacements for each dogwhistle response, we did the following: we input each of the sentences containing the replaced dogwhistle from the dataset into a sentence transformer in order to get dense 768-dimensional vector representations.

Then in order to visualize the semantic clustering of these sentence representations we used Principal Component Analysis (PCA; Abdi and Williams, 2010) to reduce the vectors to 3 dimensions.

3.3 Evaluation metrics

The general purpose of the clustering validations is to measure the compactness, i.e., how similar objects within a cluster are, and separation, which measures how far apart the clusters are. We evaluated the clustering created in the semantic space using two different evaluation metrics:

The overwhelming bulk of the training data is news media.

Davies-Bouldin (DB; Davies and Bouldin, 1979) score measures the average of the intra-cluster dispersion within each individual cluster divided by the distance between the centroid of one cluster to the centroid of the other cluster. A more compact cluster further apart from the other cluster will result in a lower score, with 0 indicating two very distinct clusters.

Calinski-Harabasz (CH; Caliński and Harabasz, 1974), measures intra-cluster dispersion and each cluster center's distance from the global centroid.

3.3.1 Unsupervised approach

We then used K-means with two cluster centroids to label each point in the space based on that point's distance from the nearest cluster centroid.

We did this with both the dimensionality-reduced sentence representations and the original 768-dimensional vectors. The sentence representations and the K-means labels were then evaluated using the aforementioned evaluation metrics.

3.3.2 Supervised approach

We evaluated the same sentence representations using the previous metrics, but with the annotated labels rather than the K-means labels. In addition, we trained a linear-kernel support vector machine (SVM). When training the SVM, we randomly sampled the sentence representations and labels, and split the data into training and testing (70%-30%). A higher F_1 score corresponds to a better division of the clusters.

4 Experiment and analysis

Our main question: is there an easily detected separation between the in-group responses and the out-group responses in the representation space?

If this was the case, it would mean that the model has picked up on some distinction between the responses that corresponds to the distinction made by the annotators. Given the distance in the semantic space between the two groups, it should be possible to separate the space with a linear SVM trained on a subset of the data.

A further question is whether there is a correlation between the clusterings and the IAA scores? Being able to linearly separate the two groups is a necessary but not sufficient condition for good clustering scores. The dogwhistle replacements might vary widely enough to not cluster well while still being separatable using a hyperplane to a high de-

Swedish	English
Flyktingpolitik	refugee policy
Berikar	enrich
Återvandring	remigration
Förortsgäng	suburban gang
Hjälpa på plats	help on location

Table 1: Swedish dogwhistles discussed in the present work and their English translations.

Dogwhistle	IAA	Responses/DWs
Flyktingpolitik	0.73/0.87	801/216
Berikar	0.79/0.91	813/102
Återvandring	0.74/0.55	776/268
Förortsgäng	1/1	816/172
Hjälpa på plats	1/0.82	788/108

Table 2: IAA for two annotation development phases and the total number of unique responses along with the subset that are in-group dogwhistle (DW) responses.

gree of accuracy. Ideally, two differentiable dense clusters would correspond to the IAA.

4.1 Results

The results in Table 3 show that a high separability among clusters does indeed correspond with the IAA agreement, which indicates the annotators ease of categorizing a response as "in-group" or "out-group". For example, the dogwhistle "remigration" had the lowest F1 score for both the dimensionality reduced sentence representations (0.72) and the original sentence representations (0.85), as well as the lowest IAA overall (0.74/0.55), as can be seen in table 2. Similarly, "suburban gang" had the highest IAA (1/1) and very high F1 scores as well (0.98/0.97).

However, the evaluation of the K-means labeled clusters did not correspond well to the IAA. The evaluation metrics for "refugee policy" is higher than "help on location" (1/0.82) despite having a much lower IAA score (0.74/0.55).

An explanation for this might be that some dogwhistle clusterings are spread over a wider semantic space, while still being linearly separatable (with an SVM) from other clusterings. This type of data distribution will still obtain good clustering results. For example, “enrich” in table 4 reports the best defined clusters overall (measured by a low DB score and high CH score), while only having a marginally greater F1 score (0.98/0.98) on the SVM task than "suburban gang" (0.98/0.97).

	3-dim	768-dim
Dogwhistle	F_1	F_1
Flyktingpolitik	0.77	0.91
Berikar	0.98	0.98
Återvandring	0.72	0.85
Förortsgäng	0.98	0.97
Hjälpa på plats	0.94	0.96

Table 3: SVM F_1 metrics for each dogwhistle.

Dogwhistle	3-dim		768-dim	
Clustering	CH	DB	CH	DB
Flyktingpolitik				
K-means	568.86	0.99	159.79	2.06
Human	65.29	2.90	40.41	3.85
Berikar				
K-means	1111.32	0.49	327.34	0.96
Human	978.04	0.61	303.33	1.12
Återvandring				
K-means	580.85	1.07	175.32	1.95
Human	148.15	2.05	64.39	3.16
Förortsgäng				
K-means	607.61	0.94	243.29	1.59
Human	241.03	1.39	115.59	2.06
Hjälpa på plats				
K-means	398.04	0.92	119.72	1.93
Human	300.58	1.02	97.16	2.02

Table 4: Cluster separability metrics for each dogwhistle for K-means and human clustering.

4.1.1 Support Vector Machine

The SVM was generally able to separate the two clusters well, even given fairly small amounts of training data. The general correlation with IAA scores were higher with PCA dimensionality-reduced vector representations. Possible reasons for the performance of the SVM might be that the SVM does not take into account the separation of the data from its cluster centroid in the opposite direction of the other cluster or the dispersion of the datapoints along an axis orthogonal to the separating plane. The SVM measurement only takes into account the overlapping of the semantic meanings of the sentences, represented in the space.

4.1.2 Internal clustering evaluation

The evaluation metrics for the K-means labeled points in the space does not seem to correspond to

the IAA values. The lowest scoring dogwhistles, "refugee policy" and "remigration", cluster fairly well compared to the other dogwhistles with higher IAA values.

4.1.3 External clustering evaluation

The results for the evaluation metrics on the human labeled points indicate that there is an overall correspondence between the IAA and those measurements: the lowest rated IAA dogwhistles always have the lowest clustering score. This indicates that there is a semantic distinction between in-group responses and out-group responses that is captured fairly well by sentence transformers.

5 Conclusions and future work

Our work contributes a computationally straightforward method to extend the manual analysis of dogwhistles that is available for many languages at a resource level similar to Swedish. Our evaluations show that easily identified dogwhistle interpretations are partitioned well enough in the vector space given by SOTA sentence models that they are linearly separable using a simple SVM.

The representation of sentences given by the model is largely derived from the corpora that the model is trained on. The corpora thus has a large impact on the semantic space. Given this, models trained on different corpora would give rise to different semantic spaces where the clustering of the sentences would be different. Since K-means does not seem to be able to differentiate between in-group sentence replacements and out-group sentence replacements, future work might include an investigation into modeling the semantic space by training a sentence transformer on different sources of text. This would also allow us to investigate the role of specific lexical choices in the detection and representation of dogwhistles. In theory, it should be possible to train a model that creates a semantic space that clusters the points in a way that that the labels can be retrieved by an algorithm like K-means using only the data itself.

Acknowledgements

Funding for this work was provided by the Gothenburg Research Initiative for Politically Emergent Systems (GRIPES) supported by the Marianne and Marcus Wallenberg Foundation grant 2019.0214. Christoffer Olssson assisted with some of the annotations used in the work.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Mathilda Åkerlund. 2021. Dog whistling far-right code words: the case of 'culture enricher'on the swedish web. *Information, Communication & Society*, pages 1–18.

Bethany Albertson. 2014. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37.

P. Ishwara Bhat and Ofra Klein. 2020. Covert hate speech: White nationalists and dog whistle communication on twitter.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Luiza Maria Filimon et al. 2020. Nordic dog whistles. analyzing discriminatory discourses in the parlance of the scandinavian radical right parties. *Revista Română de Studii Baltice şi Nordice*, 12(1):7–40.

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

Robert Henderson and Elin McCready. 2019. Dogwhistles and the at-issue/non-at-issue distinction. *Secondary Content*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden - making a swedish BERT. *CoRR*, abs/2007.01658.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Faton Rekathati. 2021. The kblab blog: Introducing a swedish sentence transformer.

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

Asad Sayeed. 2013. An opinion about opinions about opinions: subjectivity and the aggregate reader. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 691–696, Atlanta, Georgia. Association for Computational Linguistics.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Blow the dog whistle: A Chinese dataset for cant understanding with common sense and world knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2139–2145, Online. Association for Computational Linguistics.