Identifying Slurs and Lexical Hate Speech via Light-Weight Dimension Projection in Embedding Space

Sanne Hoeken, Sina Zarrieß and **Özge Alaçam**
Faculty for Linguistics and Literary Studies
University of Bielefeld, Bielefeld, Germany
{sanne.hoeken, sina.zarriess, oezge.alacam}@uni-bielefeld.de

Abstract

The prevalence of hate speech on online platforms has become a pressing concern for society, leading to increased attention towards detecting hate speech. Prior work in this area has primarily focused on identifying hate speech at the utterance level that reflects the complex nature of hate speech. In this paper, we propose a targeted and efficient approach to identifying hate speech by detecting slurs at the lexical level using contextualized word embeddings. We hypothesize that slurs have a systematically different representation than their neutral counterparts, making them identifiable through existing methods for discovering semantic dimensions in word embeddings. The results demonstrate the effectiveness of our approach in predicting slurs, confirming linguistic theory that the meaning of slurs is stable across contexts. Our robust hate dimension approach for slur identification offers a promising solution to tackle a smaller yet crucial piece of the complex puzzle of hate speech detection.[1]

1 Introduction

Recent years have seen an increase in attention towards hate speech detection due to the rising prevalence of online hate speech and its negative impact on society (Zhang and Luo, 2019). Current approaches to hate speech detection focus on identifying hate speech at the utterance level (Zampieri et al., 2020), which remains a challenging task due to the nuanced and complex nature of hate speech. Hate speech can take many different forms, especially in the context of social media platforms where language use is dynamic and constantly evolving (Davidson et al., 2017). This paper aims to tackle one aspect of hate speech detection by identifying hate speech at the lexical level, specifically through the identification of slurs based on their contextualized representations. Breaking down the problem into smaller pieces allows us to focus on specific aspects of hate speech, such as slurs, and understand how "hatefulness" is encoded as a dimension of meaning in the embedding space of language models. This, in turn, can inform the development of more robust hate speech detection methods.

Slurs can be defined as terms referencing a specific social group, and generally carry derogatory connotations, regardless of the situation in which they are used. Thus, the meanings of slurs remain relatively consistent across different contexts (Hess, 2021). In contrast, more subtle forms of hate speech such as dog whistles or expressions that depend on the speaker or audience can vary widely in their interpretation. For instance, the word "bitch" may be used as a derogatory term towards women, but among women, it can also be used casually in certain contexts (Davidson et al., 2019). Slurs are commonly included in hate speech lexicons, yet systematic study into automatically identifying them has barely been touched upon. Nevertheless, their semantic consistency across contexts makes slurs a promising target for identification based on their embedding representation.

Word embeddings have demonstrated their effectiveness in capturing various aspects of meaning, including relations like synonymy and antonymy, but also more abstract concepts like cultural or social bias. As a prominent example, previous studies have shown that gender bias can be detected by extracting the relevant semantic dimension from the embedding model (Bolukbasi et al., 2016; Garg et al., 2018). These "semantic dimension identification" techniques typically rely on a small set of carefully selected words or word pairs that only differ with respect to the semantic dimension of interest. Although this approach has demonstrated its capability to generalize to different dimensions of meaning, there are still open questions and chal-

[1] The code and accessible data used for the experiments can be found at https://github.com/SanneHoeken/SlurDimension

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 278–289
July 14, 2023 ©2023 Association for Computational Linguistics

lenges. For example, the task of selecting the right set of words or word pairs to capture a specific dimension of meaning is still indistinct.

Building on linguistic theories on slurs and the findings of dimension identification studies, we hypothesize that slurs have a systematically different representation than their neutral counterparts in the embedding space, and that this difference can be identified using existing methods for discovering semantic dimensions within word embedding models (Kozlowski et al., 2019). Specifically, our study addresses the following research questions: 1) can we identify slurs based on their contextualized word embeddings? 2) how do we leverage dimension-based methods for slur identification? 3) can we confirm existing work in linguistics which suggest that the meaning of slurs is stable across contexts? 4) can we use the hate dimension identified based on slurs for detecting other lexical units pertaining to hate speech?

In addressing these questions, we focus on methodological aspects such as the selection of lexical pairs and leveraging a pre-trained contextualized language model and incorporating multi-word expressions. In addition, our research puts emphasis on the robustness of the proposed methods across various hate speech domains and datasets without the need of big annotated data.

To sum up, this paper presents a more targeted and efficient method for detecting hate speech, that aims to identify and gain more insight into the use of slurs in online discourse. [2]

2 Related work

In this section, we review existing research on two key areas related to our study. First, we address hate speech at the lexical semantic level, with a particular focus on slurs. Second, we discuss previous work on semantic dimension identification and its applications in computational semantics.

2.1 Lexical semantics of hate speech

Hate speech can manifest itself in various forms at the lexical-semantic level, including both explicit and subtle expressions of derogatory language. Pejorative terms such as “nigger” or “faggot” fall into the former category, while more covert forms of hate speech include the use of code words and dog whistles like “inner-city” (referring to poor African-American) (Anderson and Barnes, 2022).

One prototypical (and explicit) form of hate speech is the use of slurs. Slurs are pejorative lexical items that refer to social groups defined by a factor such as race, ethnicity or religion, and convey derogatory attitudes toward those groups and their members (Hess, 2021). In his theoretical overview, Hess (2021) identifies several semantic and pragmatic properties of slurs. These properties include the observation that negative connotations of slurs persist even when used under negation, modals, or in conditionals, and that the derogatory meaning of a slur is independent of the speaker’s intentions or attitudes. This means that every use of a slur is considered offensive. Additionally, most scholars agree that for every slur, there exists a neutral counterpart that can denote the same social group without causing offense (Falbo, 2021; Bach, 2018). For example, the term “beaners” in American English is generally understood as a derogatory term used to refer to “Hispanic people” regardless of context.

The lexical aspect of hate speech has been a key focus in hate speech detection models. Earlier feature-based classification systems relied on identifying specific words and phrases that are commonly associated with hate speech, such as slurs, by employing discrete hate speech lexica (Schmidt and Wiegand, 2017). However, the explicit modeling of slurs or slur detection has not been extensively explored in this field. Currently, the only notable work in this direction is presented by Wiegand et al. (2018) who proposed a method to automatically expand a base lexicon of abusive words through a feature-based classification system. Nevertheless, their engineered features are resource-intensive as they depend on multiple corpora and lexical resources. Their system also incorporates a lexical graph propagation framework, which has been previously applied in domains beyond abusive language detection. Hamilton et al. (2016) demonstrated its applicability in generating sentiment lexicons. However, creating a lexical graph requires a semantic space that is learned from a substantial corpus of data. Furthermore, Hamilton et al. (2016) demonstrated that their method is only effective for domain-specific applications.

The linguistic properties of slurs, i.e. having neutral counterparts and invariant offensiveness, make

[2] Please note that this paper includes the use of offensive language, solely for the purpose of illustrating theoretical concepts and our proposed methodology. We acknowledge that such language may be harmful and recognize that its use does not reflect our personal beliefs or values.

them potentially suitable for a domain-independent semantic dimension approach that does not necessitate extensive data. In the following section, we will discuss the computational linguistic aspects of a semantic dimension approach in more detail.

2.2 Semantic dimensions within word embeddings

Word embedding models have demonstrated their capacity to represent shared relationships fundamental to word analogies, as constant vector offsets between pairs of words (Mikolov et al., 2013). An increasingly important line of research focuses on detecting biases with and within word embeddings. Bolukbasi et al. (2016) proposed a method based on the concept of gender direction, which involves identifying the dimension in the embedding space that captures gender information. More precisely, they take the difference vectors of 10 curated word pairs and calculate their Principal Components (PC). Subsequently, the top PC is identified as the dimension vector. Garg et al. (2018) extended this work by proposing a more general method that can identify multiple types of biases, including those related to race and religion.

Kozlowski et al. (2019) also employ the semantic dimension approach, but with a focus on analyzing cultural meaning rather than revealing bias in word embeddings. They showed that identified dimensions capturing cultural information such as affluence and status, estimated as the mean difference vector of a set of word pairs, are consistent with human-rated associations measured by contemporary and historical surveys. We adopt the dimension identification technique by Kozlowski et al. (2019) for the purpose of detecting slurs.

The analysis of semantic dimensions by leveraging the geometrical properties of the vector space has traditionally been performed using static word embedding models. Bommasani et al. (2020) introduced a novel approach to identifying social biases in pre-trained contextualized language models, such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019). The proposed technique involves reducing contextualized representations to static embeddings, which allows for the application of previously developed methods for social bias identification, as considered above.

In summary, our review of previous research has demonstrated the potential of using semantic dimension identification techniques to detect and analyze slurs, including contextualized language models. In the following section, we will describe our methodology for applying these techniques to the identification of slurs.

3 Methodology

Our approach can be divided into 1) the selection of pairs and extraction of their lexical representations from a language model, 2) the creation of a dimension vector based on those representations, and 3) the projection of lexical representations onto this dimension. We outline each of these components in Sections 3.1 through 3.3, respectively.

3.1 Extract representations

Instead of focusing on single words, we include both single words and multi-word expressions in our approach. The main motivation for this is the observation that not only slurs, but especially their neutral counterparts, are frequently encoded through multiple words. For instance, both the slur "porch monkey" and it's neutral counterpart "Black person" consist of two words. In the remainder of this paper, we therefore refer to the representations of *lexical units*, rather than words. We elaborate on our selection of lexical units in Section 4.1.

To obtain representations of lexical units from a pre-trained contextualized language model, we mainly follow techniques presented by Bommasani et al. (2020). The specifics of the particular model employed in our experiments is outlined in Section 4.3. For each lexical unit w, we sample n contexts from collected social media data that we detail in Section 4.1. We feed each context c to the model and extract the i hidden layers. Then we perform the following steps in order to compute a vector w_c for each lexical unit in each context:

1. **Layer selection & aggregation**. We average over all i hidden layers L encoding c: $L_c = mean\left(L_c^1, \ldots, L_c^i\right)$. The result is a matrix L_c of size x by d, with x being the number of subword-tokenized tokens and d the dimensionality of each layer in the model.
2. **Subword selection & aggregration**. Given all subtokens t in c, $L_c = [t_c^1, \ldots, t_c^x]$, we average over the k subword tokens generated for w_c starting at position s in c: $w_c = mean\left(t_c^s, \ldots, t_c^{s+k}\right)$. The resulting lexical representation is thus a d-dimensional vector which is the mean across all i hidden layers

and all k subtoken encodings that constitute the lexical unit w in c.

3.2 Dimension creation

Following Kozlowski et al. (2019), we calculate a semantic dimension by taking the mean of a set of pairs of lexical representations that we obtained in the previous step, whose semantic difference corresponds to the dimension of interest.

For dimension computation, we aggregate them into a single representation for each lexical unit w by taking the average representation across n contexts: $w = mean\left(w_{c_1}, \ldots, w_{c_n}\right)$. To obtain the final dimension vector v, we calculate the average difference between p pairs of static representations of a slur w^s and its neutral counterpart w^n: $v = mean\left(\left(w^s_1 - w^n_1\right), \ldots, \left(w^s_p - w^n_p\right)\right)$.

3.3 Projection onto dimension

The degree of hate encoded in the embedding of a lexical unit can be determined by its projection onto the dimension. Given an embedding w, this projection onto the dimension is defined as the cosine distance between the lexical unit and dimension vector v (Kozlowski et al., 2019). Hateful lexical units should exhibit positive projection values, and neutral terms negative values.

4 Data & Experiments

The ingredients for our semantic dimension approach using a contextualized language model are lexical units and contexts in which they occur. The subsequent sections will introduce the data that we utilize to construct dimensions (4.1) and to assess them (4.2), respectively.

4.1 Data for dimension creation

Lexical Units. As discussed earlier, slurs and their neutral counterparts seem to form ideal *pairs* of lexical units that differ only with respect to the semantic dimension of hate. To create a set of such pairs, we utilized HateBase[3], a commonly used lexicon for hate speech detection. The English Hatebase lexicon contains 1565 hate terms, including but not limited to slurs. We filtered for slurs by identifying lexical units that refer to (members of) social groups, that are hateful in any use, and for which a neutral counterpart could be found.

To find neutral counterparts, we consulted the definition of the lexical unit and the annotation of the target group as provided in HateBase, as well as definitions from resources like Wictionary[4] and other online dictionaries if needed. Furthermore, our primary objective was to comply with the APA Style guidelines for bias-free language[5].

Next, out of the resulting 617 pairs, we filtered for pairs that appeared at least 10 times in the datasets used for context sampling (which we discuss next). From this filtered list of almost 70 pairs, we selected 15 pairs in such a way that each social group was represented by no more than one word pair. Additionally, we ensured that the selected pairs provided the best possible spread across target group categories such as ethnicity and religion. Table 1 presents the final set of pairs.

	Slur	Neutral counterpart	Category
1	beaners	Hispanic people	Ethnicity
2	gooks	Asian people	Ethnicity
3	injuns	Native Americans	Ethnicity
4	Argies	Argentinians	Nationality
5	limeys	British people	Nationality
6	pakis	Pakistanis	Nationality
7	feminazis	feminists	Gender
8	tranny	transgender people	Gender
9	whore	prostitute	Gender
10	kikes	Jews	Religion
11	muzzies	Muslims	Religion
12	darkies	Black people	Race
13	whitey	White person	Race
14	hillbillies	rural people	Class
15	libtard	Liberal person	Politics

Table 1: 15 pairs of slurs and their neutral counterparts, used for dimension creation, and the category of the social group they refer to.

Contexts. In order to obtain lexical representations from a contextualized language model, we provide the model with lexical units *within contexts*. Feeding the model with isolated units (i.e. without any context) would be an unnatural input to the model. To this end, we collect a set of user-generated web-data from Reddit, a social media platform that allows users to create communities (called subreddits) based on a wide range of topics and interests. Users can submit content, such as links, text posts, images, and videos. Reddit is often used as a resource in hate speech research (e.g. Saha et al. (2019); Rieger et al. (2021)) because it has a large user base and allows for anonymity,

[3] https://hatebase.org

[4] https://www.wiktionary.org

[5] https://apastyle.apa.org/style-grammar-guidelines/bias-free-language

which can encourage people to express controversial or offensive opinions.

To obtain a diverse range of data, we utilized the Pushshift API (Baumgartner et al., 2020) to scrape a random sample of 5.8 million comments from Reddit which spans from its inception in December 2005 up to March 2023. Additionally, we collected a second dataset from Reddit that is, in contrast to the first, highly domain-specific: around 10 million comments posted in the year 2016 on The_Donald subreddit. This subreddit was created to support the United States presidential campaign of Donald Trump and was eventually banned by Reddit in 2020 for violating its policies on hate speech and inciting violence (Yurieff, 2020). By comparing the results of the two different sets of data sources (for dimension creation), we aim to gain insights into what extent the source domain affects the quality of the resulting dimension.

4.2 Evaluation data

In order to assess the representational quality of a created dimension vector we project lexical representations of a set of test terms onto the dimension, and compare the projection values with human evaluations of the hatefulness of these terms.

Lexical Units First, as a preliminary check, we established a test set that also utilizes the HateBase lexicon as data source. We selected a set of lexical units consisting of slurs and neutral terms from the nearly 70 pairs that we formed before (see Section 4.1). We selected 40 slurs and 30 neutral terms, independently, that were not part of the 15 pairs selected for dimension creation (Table 1).

Second, we test our method on a more complex task, which entails assessing terms in a context-dependent manner and allows us to draw conclusions that is not limited to the HateBase source data. To this end, we leverage the HateXplain dataset which consists of over 20,000 posts from Twitter and Gab, annotated for hate speech (Mathew et al., 2021). Notably, annotators have marked parts of the post text that could be a potential reason for its perceived hatefulness. This information is provided as the "explanation rationale" for each post, which is a list that identifies marked tokens with a 1 (denoting hateful contribution) and unmarked with a 0. We identified all the unique tokens in the HateXplain dataset, and filtered out any non-stopword nouns[6] with a frequency of more than 10. For each of the resulting 2764 terms, we collected the rationale scores assigned by multiple annotators (often 3) to each instance of the term. We aggregated the scores by taking the majority score (0 or 1) for each token. For tokens in neutral posts, for which no explanation rationales were provided, we set the scores to 0.

For our first HateXplain-sourced test set, we selected 100 nouns that refer to persons and ensured that the selection included a proportional mix of both neutral and derogatory terms. To evaluate the extent to which our slur dimension is exclusively limited to persons, we gathered a second test set that encompasses all types of nouns. Specifically, we sampled 100 nouns from the HateXplain vocabulary with an approximately uniform distribution across the corresponding average rationale scores. All three lists of test terms described here are included in Appendix A.

Contexts Each final test input includes an online post containing a particular lexical unit to provide contextualized lexical representations for all test terms. For HateXplain-sourced items, the posts available in the HateXplain dataset serve as the context. For HateBase-based lexical units that are only available without any context, we use the (general) Reddit dataset we collected in previous steps (see Section 4.1) to obtain contextualized forms of the test items. In both settings, we include the contextualized representation of *each* occurrence of a lexical unit in the test data for projection.

4.3 Experimental set-up

In our default experimental set-up, a dimension vector is computed as the mean distance vector of 15 pairs of slurs and their neutral counterpart given in Table 1. In doing so, an average lexical representation for each pair part is generated across 10 contexts taken from our collected dataset, consisting of randomly sampled Reddit comments. For the generation of lexical representations, we use the pre-trained model DistilBERT (Sanh et al., 2019), which is a distilled version of BERT and consists of 6 layers of transformer blocks, each of which has 768 hidden units.[7] Each individual contextualized representation is extracted as the average of all DistilBert's hidden layers, limited to and averaged over the sentence positions of the subwords that

[6] We employed the Natural Language ToolKit (NLTK) for selecting nouns and excluding stopwords.

[7] We implemented 'distilbert-base-uncased' through Hugging Face's *transformers* library for Python (Wolf et al., 2020).

constitute the lexical unit.

Contextualized representations of test items are projected onto a computed dimension and compared to human assessments using our three different test sets, two utilizing the HateXplain dataset and one using the HateBase lexicon (see Section 4.2). In our evaluation, we employed a combination of token-level and type-level comparison using correlation and classification metrics. To classify lexical units as hateful or neutral, we used their projection values, with positive values indicating hateful and negative values indicating neutral. For classification accuracy, we used the chi-squared test to calculate its statistical significance.

For type-level evaluation, we calculated the average projection value across all contextualized instances of each test term, whereas for token-level evaluation, we considered each instance. This allowed us to assess the performance of our method in predicting terms within their context (made possible with the HateXplain dataset) as well as utilizing HateBase, a context-independent source of hatefulness ratings. In assessing projections of HateXplain test terms, we also measured correlation using the average rationale score (0 or 1) for each term across all instances. We compared these type-level scores with the type-level projection values, using Pearson's Correlation.

5 Results

In the following, we present the results of our default set-up on two test sets, as well as experiments examining the impact of the selection of pairs (5.1). Furthermore, we analyze the effect of the number of contexts included and the domain they are sourced from (5.2). Finally, we discuss our findings on our third test set that includes other categories besides persons (5.3).

5.1 Main results

Table 2 reports our main results. Overall, our dimension approach demonstrates effectiveness in predicting slurs as evidenced by the performance results. Specifically, our method achieves accuracy rates of around 0.90 on our HateBase test set and 0,77 on the HateXplain test set limited to person terms. The higher accuracy displayed by the HateBase test set may be explained by its utilization of the same data resources as the dimension data.

Pair selection To investigate the impact of the selection of pairs on performance, we utilized two set-ups: In the first, instead of a (more specific) co-hyponym, which applies to true counterparts, we replaced all neutral counterparts with the more general hypernym "person" or "people". The semantic difference between the two pair-parts here thus involves more than purely deragotary connotation, which seems to be reflected in the resulting dimension. The second and fourth boxplot in Figure 1 show that the projection values of hateful terms in the HateXplain test set are lower overall, indicating a weaker association with the dimension. As shown in Table 2, this change caused a drop in the recall of hateful tokens in the HateXplain test set. This effect was, however, not observed for the HateBase test set. Despite observing lower projection values for hateful terms, misclassifications did not increase. One possible explanation is that the test hate terms are more similar to the slurs used for dimension creation, thereby maintaining a positive association with the dimension vector.

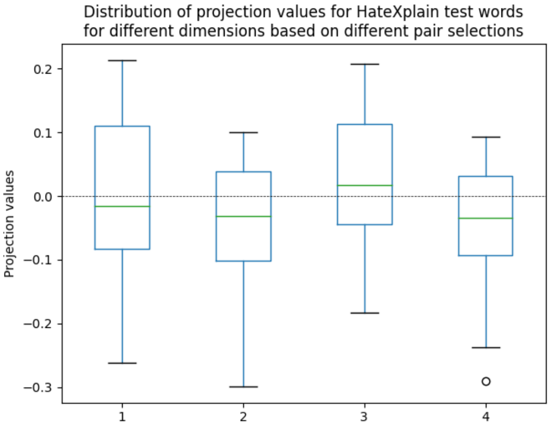

Figure 1: Effect of pair selection for dimension creation on projection values. Pair selections: *1 = all slurs & co-hyponym counterparts, 2 = all slurs & hypernym counterparts, 3 = nationality/ethnicity slurs & co-hyponym counterparts, 4 = nationality/ethnicity slurs & hypernym counterparts*

In the second set-up, we limited the set to only lexical units referring to social groups categorized by ethnicity and nationality, reducing the number of pairs to six (i.e. the first six pairs in Table 1). The resulting dimension represents a narrower spectrum of hate, which caused a significant decrease in the precision of predicting hateful terms correctly in the HateXplain data, but not in the HateBase test set. This discrepancy may be due to the majority of HateBase test terms referencing nationality or ethnicity categorized groups. Additionally, we found that combining hypernym counterparts with a restrictive set of slurs did not result in an increase

Pairs		Correlation	Classification report						
Slurs	Counter parts	Pearson's r (* = sig.)	Acc. (* = sig.)	Hateful			Neutral		
				Prec.	Recall	F1	Prec.	Recall	F1
HateBase test set									
All	Co-hyponyms	-	0.90*	1.00	0.82	0.90	0.81	1.00	0.90
All	Hypernyms	-	0.89*	0.97	0.82	0.89	0.81	0.97	0.88
Nat./Eth.	Co-hyponyms	-	0.91*	1.00	0.85	0.92	0.83	1.00	0.91
Nat./Eth.	Hypernyms	-	0.89*	0.97	0.82	0.89	0.81	0.97	0.88
HateXplain test set - persons									
All	Co-hyponyms	0.790*	0.77*	0.76	0.86	0.81	0.79	0.66	0.72
All	Hypernyms	0.755*	0.77*	0.80	0.80	0.80	0.74	0.74	0.74
Nat./Eth.	Co-hyponyms	0.770*	0.76*	0.71	0.95	0.82	0.90	0.51	0.65
Nat./Eth.	Hypernyms	0.737*	0.77*	0.80	0.79	0.80	0.74	0.75	0.74

Table 2: Performance results for dimensions with different pair selections and different test sets. (n = 15 for *All* slurs, n = 6 and for *Nat./Eth.* (Nationality/Ethnicity) slurs)

in false positives in either test set. This could be because the effect of hypernym counterparts in decreasing false positives is stronger.

Overall, these findings underscore that manipulating the hate specificity of the dimension by selecting different pairs does not significantly impact the overall accuracy and F1-scores. However, it does have a notable effect on the occurrence of false negatives or false positives, which is particularly relevant for hate speech detection.

5.2 Number & domain of contexts

Number of contexts Each lexical representation is produced based on the 10 contextual representations (as mentioned in Section 4.3). To evaluate the necessity of such data quantity, we conducted projection tests with dimensions based on less than 10 contextualized representations per lexical unit.

Figure 2 depicts the impact of the number of contexts on dimension performance, with detailed results presented in Table 6 in Appendix B. The results indicate that larger sample sizes result in greater stability in performance, as evidenced by reduced variation introduced by random sampling. Yet more importantly, our analysis suggests that the effectiveness of a dimension is not significantly influenced by the size of the context sample. This implies that accurate results could be obtained even with smaller amounts of data, thereby providing a more efficient and cost-effective method.

Domain of contexts In addition to the quantity of contexts, we also tested the influence of the domain from which the contexts were sourced. Rather than sampling contexts from comments across the entire Reddit spectrum, we conducted an additional experiment using domain-specific comments from The_Donald subreddit, specifically those from 2016. This change did not reveal any significant effect on the dimension performance, as evidenced by the performance results in Table 3. These findings align with our previous results regarding sample size, and furthermore, they indicate that domain-specific data is also appropriate for our method, thereby increasing its versatility. Most importantly, the results confirm linguistic theory that the meaning of slurs is stable across contexts (Hess, 2021).

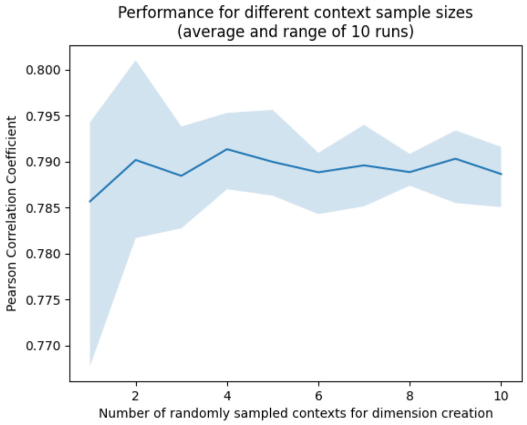

Figure 2: Effect of number of contexts on performance on HateXplain test words

5.3 Generalizing to other categories

In our final experiment, we tested our dimension approach on another dataset comprising 100 random nouns from the HateXplain vocabulary in their respective contexts. We observe a significant drop in

Correlation	Classification report						
Pearson's r	Acc.		Hateful			Neutral	
(* = sig.)	(* = sig.)	Prec.	Recall	F1	Prec.	Recall	F1
0.790*	0.77*	0.76	0.87	0.81	0.79	0.65	0.71

Table 3: Performance results for a dimension based on domain-specific contexts on HateXplain test words

performance when predicting nouns that were not restricted to persons, like "stupidity" and "lottery". In particular, the correlation coefficient decreased by 0.3 (See Table 4).

Correlation	Classification report						
Pearson's r	Acc.		Hateful			Neutral	
(* = sig.)	(* = sig.)	Prec.	Recall	F1	Prec.	Recall	F1
0.482*	0.55*	0.43	0.53	0.47	0.66	0.56	0.61

Table 4: Performance results for dimension on HateXplain test set comprising random nouns

An analysis of the projections of the non-persons test terms (on token level) shows that most errors were false positives, with a significant number of neutral terms incorrectly predicted as hateful (See confusion matrix in Table 5.

		Projection	
		Hateful	Neutral
Gold	Hateful	1487	1337
	Neutral	1992	2562

Table 5: Confusion matrix for classification on HateXplain test set comprising random nouns

Many of these false positives were non-nouns, like "pay", "lmao" and "pro". This shows some inadequacies of the noun filtering method for the construction of the test set, as well as the need for greater part-of-speech robustness. Other false positives comprised nouns that did not refer to persons, such as "tweets" and "propaganda". Nonetheless, our methodology also demonstrated its ability to correctly label such terms, as evidenced by the correct prediction of e.g. "movement", "prison", and "knowledge".

Our analysis of false negatives has revealed limitations in using the HateXplain dataset as gold data for our specific purpose. We attribute this issue to the distinction between utterance-level and lexical-level purposes, which we have touched upon in the introduction. The human rationale scores in the HateXplain dataset reflect a word's contribution to the overall hateful meaning of the utterance. We, on the other hand, employed them as evaluations of the hatefulness of a specific lexical unit within a given context. This approach posed problems as demonstrated by the largest group of false negatives, which include terms that reference target groups but do not necessarily contain derogatory connotations at a lexical-semantic level, such as "feminist", "homosexuals" and "refugee".

Lastly, the results also indicated promising classification beyond the intended slur detection: Firstly, the method detected a hateful term that does not refer to persons, i.e. "holohoax". Secondly, the method detected the ambiguous term "fruit", that appeared to be used derogatorily to refer to LGBT people in certain contexts. For example: " *yep and he meets that satanic **fruit** every week how r... and g... is this man*". These findings suggest that our method has a potential wider application in detecting offensive language beyond just slurs.

6 Discussion & Future directions

The results indicate that our dimension approach is effective in predicting slurs based on their contextualized embedding, with the importance of selecting pairs carefully to create a robust hate dimension. Due to the lack of a universally agreed-upon definition of hate speech, the creation of hate speech datasets is difficult and prone to bias (Davidson et al., 2017; Waseem and Hovy, 2016). As a result, datasets are often limited in size and scope, making it challenging to train models that can effectively detect a wide range of hate speech in different domains. Our results demonstrated that the effectiveness of the dimension is not significantly influenced by the size of the context sample. This implies that our dimension approach is a promising cost-effective and domain-agnostic method for identifying slurs with low-data requirements.

6.1 Generalizability

When it comes to classifying *non-person* nouns, we observe a decline in the performance. However, our approach also shows promising results in detecting other categories than slurs, opening a possibility for extension beyond slurs. The further analysis indicated that many errors can be attributed to the quality of evaluation data rather than inherent limitations of the method itself (Section 5.3). Regarding the data employed in our study, we have selected a diverse yet bounded domain coverage, for the purpose of maintaining a systematic approach. Nevertheless, it is worth noting that our findings encourage further exploration of performance in alternative contexts. To illustrate this point, it would be interesting to observe how our method performs when faced with phenomena such as the non-derogatory use of the n-word slur within certain in-group contexts.

6.2 Technical considerations

Prior research on extracting lexical representations from models like BERT demonstrated significant effects of hidden layer selection on the efficacy of the derived representations for various lexical-semantic tasks (Vulić et al., 2020; Bommasani et al., 2020). While averaging all hidden layers generally yields beneficial representations, no single layer configuration stands out as the overall best. The optimal configuration appears to depend heavily on the task and methodology employed. Future research should investigate alternative layer configurations to improve the effectiveness of the representations for identifying slurs.

Moreover, it is important to experiment with different definitions of dimension computation in future research, such as PCA-based (Bolukbasi et al., 2016) or vector offset-based methods (Garg et al., 2018). This is particularly crucial since Bommasani et al. (2020) demonstrated the significant effect of the bias quantification method on the measured bias in lexical representations.

7 Conclusion

This paper addresses the complex puzzle of hate speech detection by breaking it down and concentrate on a smaller but crucial piece, the identification of slurs. We propose a novel approach that applies semantic dimension identification with contextualized embeddings to the detection of slurs. In this study, we set out to address several key research questions concerning the identification of slurs. First, we investigated whether slurs can be identified based on their contextualized word embeddings. Our experimental results demonstrated the effectiveness of our method in predicting slurs by leveraging contextual representations, thereby affirming their effectiveness. Simultaneously, we explored the application of dimension-based methods for slur identification. Our findings highlight the significance of carefully selecting lexical pairs while demonstrating that extensive data is not necessarily required. Additionally, we aimed to confirm existing work in linguistics, which suggests that the meaning of slurs is stable across contexts. Findings on our experiments across different domains and datasets align with linguistic theory, as evidenced by consistently strong prediction performance. Lastly, we explored the potential of utilizing the hate dimension identified based on slurs for detecting other lexical units related to hate speech. Our method exhibited promising results in detecting other categories of lexical hate speech, showcasing its broader applicability potential beyond slurs. In conclusion, our approach contributes to a more targeted and efficient method for detecting hate speech and sheds light on the use of slurs in online discourse.

Acknowledgements

The authors acknowledge financial support by the project “SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), which is funded by the program “Netzwerke 2021" of the Ministry of Culture and Science of the State of Northrhine Westphalia, Germany.

References

Luvell Anderson and Michael Barnes. 2022. Hate Speech. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2022 edition. Metaphysics Research Lab, Stanford University.

Kent Bach. 2018. Loaded words: On the semantics and pragmatics of slurs. In *Bad Words: Philosophical Perspectives on Slurs*, pages 60–76. Oxford University Press Oxford.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Arianna Falbo. 2021. Slurs, neutral counterparts, and what you could have said. *Analytic Philosophy*, 62(4):359–375.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Leopold Hess. 2021. *Slurs: Semantic and Pragmatic Theories of Meaning*, Cambridge Handbooks in Language and Linguistics, page 450–466. Cambridge University Press.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and reddit. *Social Media+ Society*, 7(4):20563051211052906.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*, page 255–264. Association for Computing Machinery.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kaya Yurieff. 2020. Reddit bans pro-trump forum the_donald and other communities that promote hate. *CNN Business*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semant. Web*, 10(5):925–945.

A Lists of lexical units for evaluation

HateBase: *african americans , albanians, american indian, americans, anchor babies, arabs, armenians, australians, azn, bimbo, bint, canadians, chav, chinaman, chinese people, ching chong, cunt, dutch people, faggy, filipinos, french people, gay people, gaylord, german people, ghey, goatfucker, gypsy, honky, hoodrat, immigrant, indian people, inuit, irish people, italian women, leb, mexican immigrants, middle-class people, mongoloid, native hawaiian, newfies, nigger, oklahoman, pacific islanders, palestinian, polack, polish people, porch monkey, protestants, race traitor, redneck, refugees, roman catholics, scally, seppo, shemale, shyster, slut, spics, sub human, taqiyya, trailer trash, twat, waspy, wetback, white trash, wigger, women, yokel, zio, zog*

HateXplain - persons: *asshole, bigot, bitch, boomer, boyfriend, brother, buddy, captain, chinaman, citizens, clown, cocksucker, commies, coons, coward, cuckservatives, cunts, doctors, driver, dykes, faggot, farmers, fascist, followers, friends, fuckers, girls, goatfucker, governor, haters, heeb, hero, hoes, honky, idiot, jigaboo, jockey, journalists, kids, ladies, lawyer, leader, leftie, loser, manager, moron, mother, mudshark, mudslime, muzrat, negress, negros, nigger, officers, officials, parents, partners, patriots, pedos, politician, prayers, president, princess, professor, protesters, pussy, queen, racists, raghead, rapefugees, rapper, redneck, residents, retard, sandniggers, satan, satanist, savages, scumbag, sheboon, sheriff, shitlib, shitskin, sjw, slave, slut, spics, students, taxpayers, teachers, towelhead, traitors, twat, veterans, warriors, wetbacks, whore, wigger, workers, yid*

HateXplain - random nouns: *action, aids, aliens, ape, army, ass, banislam, bat, beaners, bitch, bread, brown, charge, chinaman, code, commit, crack, cum, degeneracy, dicks, dislike, dumbass, faggotry, feminist, filth, friday, fruit, fuckers, gap, ghetto, girls, goatfucker, goy, head, hebrew, holohoax, homophobic, homosexuals, husband, illegals, infidels, jewish, khan, knowledge, lit, lmao, lottery, mans, mexicans, monkey, moslem, movement, mudslimes, muslime, muzrat, muzrats, nazi, negress, nig, niglet, noise, paki, pakis, pay, pedophile, pedophiles, pedophilia, players, porch, posts, prayer, prison, pro, propaganda, rag, raghead, rapist, redneck, refugee, ricky, savages, sheboon, shitskin, socialists, sort, steal, stupidity, subhuman, subversive, thot, thots, thru, trans, tweets, values, weird, wetbacks, wigger, witch, yid*

B Additional results of experiments

Sample size	Correlation	Classification report						
	Pearson's r (* = sig.)	Acc. (* = sig.)	Hateful			Neutral		
			Prec.	Recall	F1	Prec.	Recall	F1
1	0.786*	0.76*	0.76	0.86	0.80	0.78	0.64	0.70
2	0.790*	0.77*	0.76	0.86	0.80	0.78	0.65	0.71
3	0.788*	0.77*	0.76	0.86	0.81	0.78	0.65	0.71
4	0.791*	0.77*	0.76	0.86	0.81	0.79	0.65	0.71
5	0.790*	0.77*	0.76	0.87	0.81	0.79	0.65	0.71
6	0.789*	0.77*	0.76	0.86	0.81	0.78	0.66	0.71
7	0.790*	0.77*	0.76	0.85	0.81	0.78	0.66	0.72
8	0.789*	0.77*	0.76	0.86	0.81	0.79	0.65	0.71
9	0.790*	0.77*	0.77	0.85	0.81	0.78	0.66	0.72
10	0.789*	0.77*	0.76	0.86	0.81	0.78	0.66	0.72

Table 6: Performance results (average over 10 runs) for dimensions with different context sample sizes