FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and Information Science

Offensive language exploratory analysis

Maša Kljun, Matija Teršek

Abstract

In this paper we focus on the exploratory analysis of 10 different subgroups of hate speech. We use natural language processing techniques in order to find the underlying structure and connections/relations between the subgroups. We focus on data extracted from Twitter and online forums. First we use classic approaches, such as TF-IDF, BoW, and LDA, then we move on to more sophisticated methods such as embeddings. We use both non-contextual embeddings, such as Word2Vec and GloVe, and contextual embeddings, such as BERT. We find out that TODO: Describe the findings for the last submission.

Keywords

Hate speech, TF-IDF, embeddings, exploratory analysis, NLP ...

Advisors: Slavko Žitnik

Introduction

In the last few years social media grew exponentially and with it also the ability of people to express themselves online. By enabling people to write on different online platforms without even identifying themselves it lead to a new era of freedom of speech. As this new medium for communication and writing brought many positive things, it also has its downside. Social media has become a place where heated discussions happen and often result in insults and hatred. It is an important task to recognize hate speech and to prevent it.

Hate speech is defined as *abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation*[1]. We can see that the definiton is very vague. Having said that, the goal of our paper is to help distinguish different types of hate speech and find the specific keywords of its subgroups in order to explain its structure. This could help with its identification and classification. In this paper we focus on ten subgroups of hate speech - *abusive, hateful, spam, general hate speech, profane, offensive, cyberbullying, racism, sexism,* and *benevolent sexism*. With understanding the structure of these groups, the goal is to also find similarities and connections between them.

There has been done a lot of research regarding the hate speech, however these works are usually focused on the classification of hate speech. One of the first works include [2] who built the decision tree based classifier Smokey for abusive message recogniton and classification. Some other works that focus mainly on classification include [3] who compare the classification accuracy of models trained on expert and amateur annotations, [4] who use convolutional neural networks for classification into four predefined categories, and [5] who use different natural language processing techniques for expanding datasets with emotional information for better classification. In the last years, especially deep learning models are often used for detection and classification of hate speech, such as [6] who propose a sophisticated method that is a combination of a deep neural network architecture with transfer learning. There is a also a lot of related work that focuses on creating large datasets such as [7] who create a large-scale, multilingual, expert based dataset of hate speech.

What is less common in the research area of hate speech is analysis of relationships between different types of hate speech and the importance of specific keywords. Some examples include [8], who try to separate bullying from other social media posts and try to discover topic of bullying using topic modeling with Latent Dirichlet Allocation (LDA). [9] model hate speech against immigrants on Twitter in Spain. They try to find underlying topic of hate speech using LDA, discovering features of different dimensions of hatespeech, including foul language, humiliation, irony, etc. [10] conduct a survey about hate speech detection and describe key areas that have been explored, regarding the topic modeling, as well as sentiment analysis.

This paper is organized as follows: we present the datasets of tweets and comments in Section 1, we present our data preprocessing routine in Section 2, we perform the exploratory analysis by using many traditional and neural approaches in

Section 2, and we show the final results and a scheme of hate speech in todo

1. Data

We use six publicly available datasets for our exploratory analysis. We combine datasets [3], [6], and [11] into one large dataset (reffered to as Dataset SRB) as they include same categories of hate speech. We make labels *sexism*, *racism*, and *both* from [3] and [6]. The third dataset ([11]) that we use contains label *hostile sexism*, where marked tweets are already included in the first two datasets under *sexism*, and label *benevolent sexism*, which we rename to *benevolent*. We obtain a dataset with 6069 samples that are labeled either *sexism*, *racism*, *both*, or *benevolent*. The fourth dataset (reffered to as Dataset AHS)[12] that we use has 3 categories - *abusive, hateful, spam*. As this is the original dataset no additional merging is needed. We obtain a dataset with 13776 tweets with the mentioned labels. Note that we exclude *None* label from both datasets, as we do not need it for the analysis. We show the distribution of individual categories from datasets SRB and AHS in Figures 1 and 2, respectively. Note that the numbers of samples might not match the numbers in the original papers, due to the Twitter removing the tweets, making them unavailable for us to analyze. We also provide an example for each label.

Racism - "He can't be a server at our restaurant, that beard makes him look like a terrorist." Everyone laughs. #fuckthanksgiving

Sexism - #katieandnikki stop calling yourselves pretty and hot..you're not and saying it a million times doesn't make you either...STFU

Benevolent - It's "NEXT to every successful man, there's a woman"

Spam - RT @OnlyLookAtMino: [!!] #WINNER trending #1 on melon search

Abusive - You Worried About Somebody Bein Ugly... Bitch You Ugly...

Hateful - i hope leaders just kick retards that fake leave teams today

Additionally we use the dataset of comments extracted from the League of Legends community [13]. We preprocess the dataset given in the SQL format to a more readable CSV form and keep only the posts that are annotated as harassment. We obtain 259 examples of cyberbullying examples. The sixth dataset that we use was designed for the problem of the hate speech identification and classification, but we use the labels from the train and test set and merge them into one big dataset that we use for our analysis. It provides tags of *hatespeech*, *profane*, and *offensive*, so we refer to the dataset as HPO. It

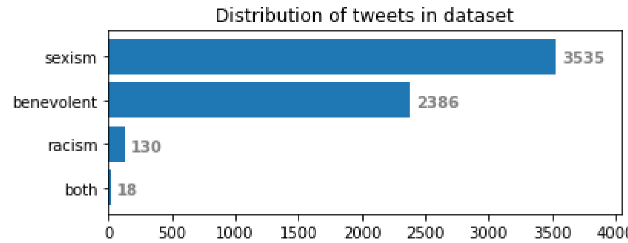

Figure 1. Distribution of tweets in SRB dataset. This figure shows the distribution of hate speech categories in the SRB dataset. We can see that *sexism* and *benevolent* are well represented, whereas *racism* and *both* are far less frequent. Original set contains more tweets labeled *racism*, but due to their removal we cannot obtain them.

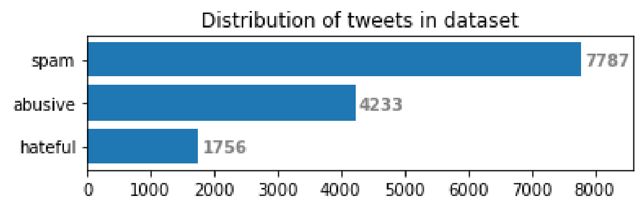

Figure 2. Distribution of tweets in AHS dataset. We see that the *spam* is the most represented label in the dataset, which represents the majority of the dataset. This is followed by the *abusive* tweets and there is the least *hateful* tweets. We can see that categories in this dataset are well represented.

consists of 2549 tweets, distribution of which can be seen in Figure 3. We again provide an example for each of the labels.

Cyberbullying - plot twist she's a fggt

Hatespeech - Johnson you liar. You don't give a flying one for the Irish

Offensive - #FuckTrump And retired porn star Melania too.

Profane - Fuck Trump and anybody who voted for that Lyin POS! #FuckTrump

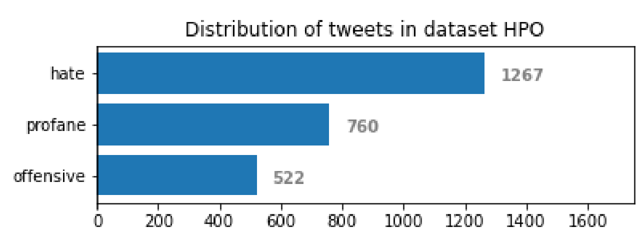

Figure 3. Distribution of tweets in HPO dataset. The most used label is *hatespeech*. It is followed by *profane* and then *offensive*, which have a similar number of tweets.

We also use the dataset of Wikipedia comments [14], that are marked as either *toxic*, *sever toxic*, *obscene*, *identity hate*, *threat*, and *insult*. We merge the first two categories into *toxic*.

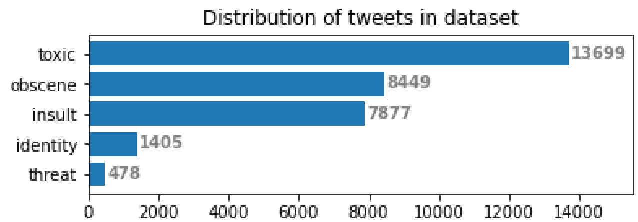

Figure 4. Distribution of tweets in TOITI. We see that most of the comments are labeled as *toxic*. Around half of them are *obscene* and around half are also labeled as *insult*. *Identity hate* and *threat* are far more uncommon in this dataset.

It is important to note that each comment in this dataset might have multiple labels, so the results for those tags might be similar. Original dataset contains 159571 tweets, 16225 of which are labeled. We show the distribution of the labels in Figure 4. We denote this dataset as TOITI in the future text.

Threat - SHUT UP, YOU FAT POOP, OR I WILL KICK YOUR ASS!!!

Obscene - you are a stupid fuck and your mother's cunt stinks

Insult - Fuck you, block me, you faggot pussy!

Toxic - What a motherfucking piece of crap those fuckheads for blocking us!

Identity - A pair of jew-hating weiner nazi schmucks.

2. Data preprocessing

Before applying any methods we first preprocess all of our data. We separate datasets into subgroups only, where each contains multiple documents - tweets belonging to this category. We remove retweet text RT, hyperlinks, hashtags, taggings, new lines, and zero length tweets. We further filter out tokens that not contain letters, e.g., raw punctuation.

Methodology

We start the analysis with more traditional approaches, and continue with neural approaches.

LDA

We use Latent Dirchilet Allocation (LDA) in combination with Bag-of-Words (BoW) and TF-IDF in hopes of finding obvious topics from all the provided comments / tweets. We try to determine 15 different topics, which is the same as the number of labels we have in our datasets. Results using BoW and TF-IDF are similar, however, we cannot clearly distinguish between the topics and connect obtained topics to the existing labels, aside from one topic, which is related to sexism. Top 5 most related words are: *penis, rape, image, live, vagina*.

TF-IDF

We continue with the analysis of datasets with a traditional method TF-IDF as we want to see the most relevant words for each category of offensive language that we have in the dataset. We show the results in Table 1. We can see that some of the categories have similar unigrams that achieved the highest TF-IDF score. An example of categories with the same highest scored unigrams are *insult* and *obscene*. This makes it harder to differentiate between the categories. It is important to note, that such examples might also occur due to subjective labeling in the provided datasets, as well as people not clearly differentiating between these categories. Most datasets are not labeled by experts, but with the help of platforms such as FigureEight or Amazon Mechanical Turk. From the results in Table 1, we could assume that most people perceive categories such as *insult* and *obscene* or *threat* and *toxic* similarly. On the other hand, categories such as *spam* or *cyberbullying* are clearly differentiable from other categories. We can also see a lot of categories including Trump related words (*hatespeech, profane*, and *offensive*). Those categories are taken from the same dataset, and we can see that such labels will contain words that are related. So the words connected to those labels might also be connected to some bigger topic, which depends on the annotator's choice from where to extract the tweets / comments.

category	**unigrams with highest TF-IDF score**
racism	peopl, white, terror, man, look
sexism	feminazi, women, think, sexist, notsexist
benevolent	women, classi, sassi, nasti, gonna
abusive	know, stupid, shit, like, idiot
hateful	peopl, trump, nigga, like, idiot
spam	giveaway, game, enter, work, home
cyberbullying	one, guy, good, gone, go
hatespeech	world, trumpisatraitor, trump, shameonicc, peopl
identity hate	fuck, shit, littl, like, one
insult	delet, go, ass, stupid, bitch
obscene	delet, go, stupid, bitch, ass
offensive	trumpisatraitor, like, douchebag, fucktrump, get
profane	trump, shit, say, resist, peopl
threat	fuck, get, die, want, find
toxic	fuck, get, bitch, want, block

Table 1. Table shows 5 highest scoring unigrams for each label we investigate. We choose the parameters, which we believe provide us the most meaningful unigrams, so we consider words that appear in at least 5% and less than 60% of the documents.

Non-contextual word embeddings

We first find most similar words to the category label names and use the obtained results to find similarities between labels. For this task we use pre-fitted Word2Vec ([15], [16]), GloVe [17], and FastText ([18]). We visualize the results with the help of t-SNE. Because of this we cannot interpret distances between the labels from the visualization. However, we can still infer that the labels that are intertwined are more similar than those that are nicely separable from one another.

We show the results in Figure 5. We can see that *homophobic* and *racist* appear very intertwined in Word2Vec and

GloVe embeddings, meaning that they cannot be separated, thus indicating a strong relation. On the other hand, in both of these embeddings *spam*, *toxic*, and *discredit* are well separated from other groups and are clearly distinguishable from others. We can also see that *abusive* is entangled with *benevolent* in GloVe representation, however, in results obtained from Word2Vec *benevolent* is nicely separable from other labels. So it is difficult to conclude that *benevolent* is a label that is different enough from other labels. FastText also nicely separates *toxic* and *benevolent* from other labels, but is unable to separate *vulgar*, *profane* and *obscene*, and *insult*. From all three models combined, we can conclude that the only label that can be always well distinguished from the others is *toxic*, and that *vulgar*, *profane*, *obscene*, and *insult* are labels that cannot be nicely separated. From all three models we can also see that *discredit* and *insult* are slightly intertwined, indicating some sort of weaker relationship. We also conclude that *spam* is a nicely separable category. Note that in some models we omit labels that are not in a vocabulary.

By now we provide some relations and decide to further investigate the connections between the related labels using word analogy. We try to find hyponyms and hypernyms, which we do with the help of the following setting:

```
father : son = our_label : x   (hyponyms)
animal : cat = our_label : x   (hyponyms)
son : father = our_label : x   (hypernyms)
cat : animal = our_label : x   (hypernyms)
```

where `our_label` is one of the analyzed labels and `x` is the word found by Word2Vec or Glove.

Unfortunately, the relationships are not clear and uniquely defined. An example is *racism* is to *sexism* what is `son` to `father` with $\approx 64.6\%$ probability, but *sexism* is to *racism* what is `son` to `father` with $\approx 64.8\%$ probability. We can once again see that the two labels are related, but the precise relationship cannot be inferred. Using `brother` and `sister` the probability is lower. This could indicate that it is impossible to find a specific hypernym and that we can only conclude that the labels are more closely related to each other, as they are each in some way hypernym and hyponym of each other. Similarly, *racism* and *sexism* are connected to *homophobia* and *slur*. Another group that we find, but also cannot clearly define the inner relations contains *vulgar*, *profane*, and *obscene*.

As mentioned, the distances between the inspected labels cannot be determined from our chosen visualization. That is why we approach this problem with clustering. We use k-means and hierarchical clustering in hopes of finding meaningful clusters that could help us understand the relationships between the subgroups of the hate speech better. We determine the k in k-means by using the silhouette score. Note that we choose the k of the second peak of the score, as we try to better divide the labels into subcategories than just two. See the example output of the silhouette score in Figure 6.

From the top 30 similar words for each label, we compute an average vector and we obtain one such vector for each label.

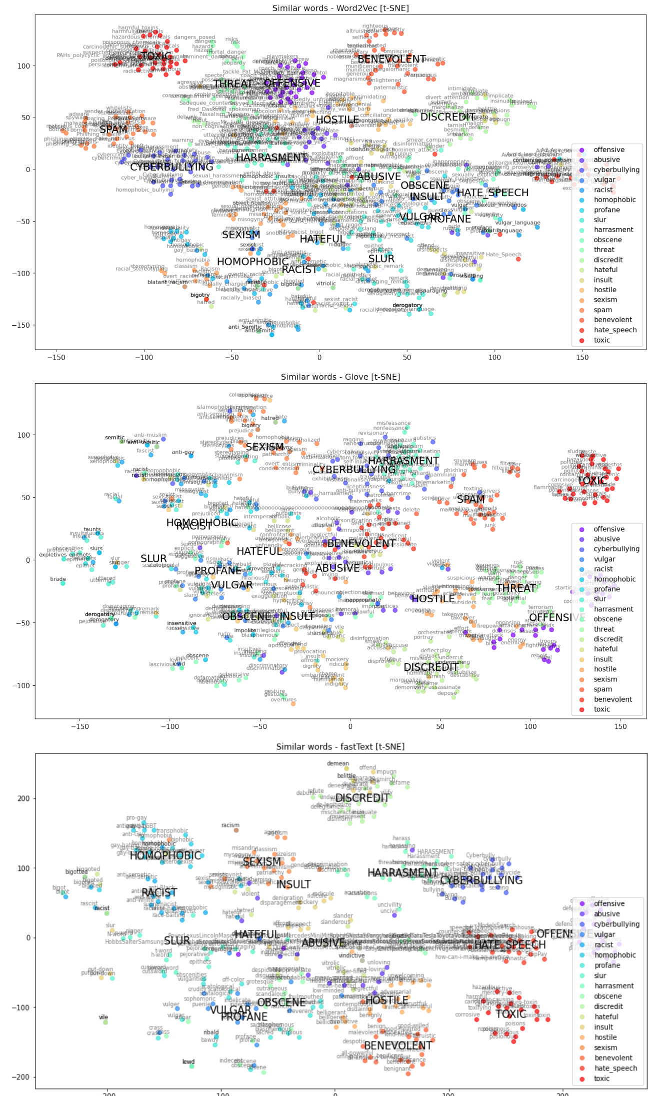

Figure 5. Word2Vec, Glove, and FastText similar labels. Figure shows Word2Vec (1st row), Glove (2nd row), and FastText (3rd row) embeddings of neighboring words of labels we analyze. Note that we omit labels that are not in vocabulary.

We compute the cosine similarity matrix between the vectors *simcos* and compute the distance matrix as $d = 1 - simcos$, which we then use for the clustering. In Table 2 we show the obtained clusters and in Figure 7 we show the results of hierarchical clustering of Word2Vec embeddings.

From these two clustering results we can infer that *insult* and *obscene* are two similar subgroups of hate speech as they both appear in the same cluster in+ k-means clustering and we can see that they are closely together in hierarchical clustering. They are also very similar according to the results from TF-IDF as seen before. *Benevolent sexism* is also close. We can see that *cyberbullying* and *spam* are clustered together in both clusterings and that *threat* and *toxic* are also very similar. From the results of hierarchical clustering we can see that

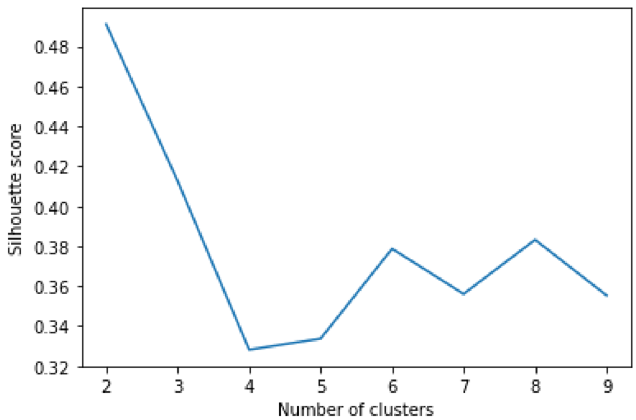

Figure 6. Silhouette score. Example of silhouette scores for different numbers of clusters. We use the second peak ($k = 6$) instead of first ($k = 2$), as we want to get more clusters.

offensive is also close to them.

Comparing the hierarchical clustering results of Glove and FastText embeddings to Word2Vec embeddings, we can see that we always get almost the same two main clusters as those in Figure 7, so we do not show figures with those results.

Looking at k-means clustering of Word2Vec and GloVe embeddings we see that pairs of *abusive*, *vulgar*, *racist*, *homophobic*, *profane*, *slur*, *obscene*, *hateful* and *insult*, and *discredit* and *hostile* always appear in the same two clusters, so we can conclude that they are related. We do not include the results of FastText k-means clustering, as its silhouette score is ≤ 0.30 for all possible k, whereas in the first two, the score is often > 0.30.

We try to apply this same approach to the words with highest TF-IDF scores from each subgroup, however, the obtained clusters provide no useful understanding, so we omit those results.

cluster	components
1	offensive
2	abusive, vulgar, racist, homophobic, profane, slur, harrasment, obscene, hateful, insult, sexism, hate speech
3	discredit, hostile, benevolent
4	cyberbullying, spam
5	threat, toxic

Table 2. K-means clustering of average Word2Vec embeddings of labels' 30 nearest words. Table shows five clusters obtained with 5-means clustering. We determine $k = 5$ using silhouette score.

Contextual word embeddings

We move on to contextual embeddings and we focus on BERT. We use the pretrained BERT base cased model [?] and convert tweets and comments from our dataset to BERT embeddings.

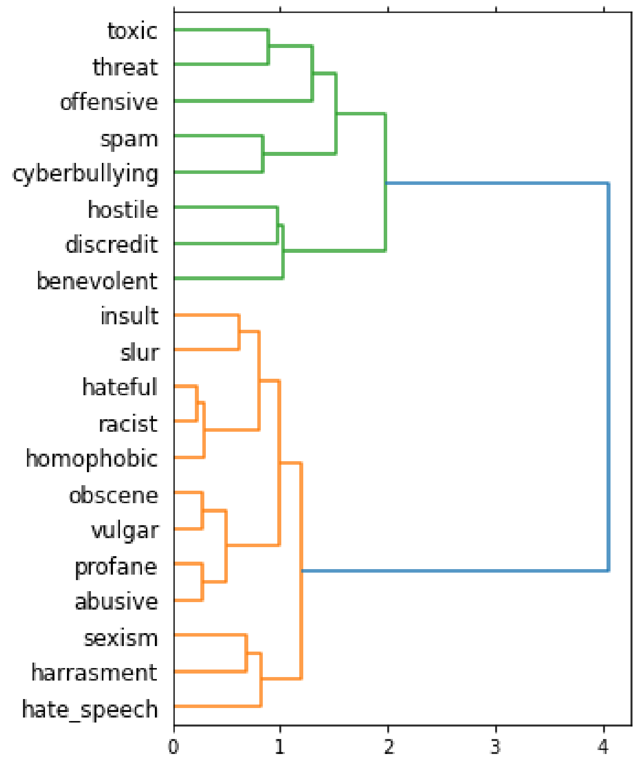

Figure 7. Hierarchical clustering of average Word2Vec embeddings of labels' 30 nearest words.

We first append them – `This is <label>` and compute the embeddings. From obtained embeddings of each vector, we compute an average representation from the vectors that belong to the tokens of the label. We average the obtained representation of each label and use cosine similarity to compute the similarity between those label representations. We show the obtained similarity matrix in Figure 8. We can see high similarities between most of the subgroups of hate speech. The one that differs the most from the other groups is *cyberbullying*. We can also see that *profane* is slightly less similar to *identity*, *insult*, *threat*, and *toxic*, however the similarity score is still between 0.87 and 0.89. For all other combinations the similarity score is ≥ 0.90.

Using sentence transformers [19] we obtain embeddings for each tweet/comment without the `This is <label>` appendix. For each label we try to find the keywords that describe subgroups of hate speech the most. We do this by first embedding the tweets/comments using BERT and separately embedding the sub-phrases from documents. For each document we then try to find the most similar sub-phrases with the help of embeddings and cosine similarity. For each label we provide three of the most common keywords obtained from the documents. We show the keywords in Table 3.

Discussion

In the last submission we are going to discuss the findings and provide the schema of hate speech.

category	BERT keywords
racism	racistreading, terroristreligion, terroristnationpakistan, faithlessfaggotboy, racisttrying
sexism	godlesswomen, sexlifeless, bitchmattythewhite, onlyallwomen, bitchesvagina
benevolent	sexyolderwomen, girlsbeautiful, bitchmattythewhite, womenattractmenattracted, happywomenday
abusive*	misogynistic, hatemyowngender, hatefultrump, selfishmotherfuckers, faithlessfaggotboy
hateful*	misogynistic, hatemyowngender, selfishmotherfuckers, terroristhasreligion, faithlessfaggotboy
spam	100million, enjoyed, enjoying, fridaymotiovation, saturdaynightonline
cyberbullying*	whoremonger, selfishmotherfuckers, faithlessfaggotboy, bitchmattythewhite, boycottfoxnews
hatespeech*	doctors_against_assualt, trumpobstructedjustice, trumpfascism, terroristnationpakistan, doctorsfightback
identity hate	whoremonger, gaywad, nazisnotwelcome, racisttrying, racistreading
insult*	whoremonger, selfishmotherfuckers, boycottfoxnews, killyourself, cocksuckerfuck
obscene*	boycottfoxnews, whoremonger, killyourself, selfishmotherfuckers, cocksuckerfuck
offensive*	trumpobstructedjustice, trumpfascism, selfishmotherfuckers, murdermystery, terroristhasreligion
profane*	trumpfascism, trumpobstructedjustice, faithlessfaggotboy, selfishmotherfuckers, doctors_against_assualt
threat	killyourself, murdering, whoremonger, hatemongers, murdermystery
toxic*	boycottfoxnews, selfishmotherfuckers, killyourself, whoremonger, nazisnotwelcome

Table 3. BERT keywords. Table shows 5 most important keywords for each hate speech subgroup found with BERT. Note: categories that have an contain also 2 keywords (dumbdonnythedraftdodgingdotart and worstsecretaryofstateinushistory) but we omit them in order to obtain clearer represenatation of categories.

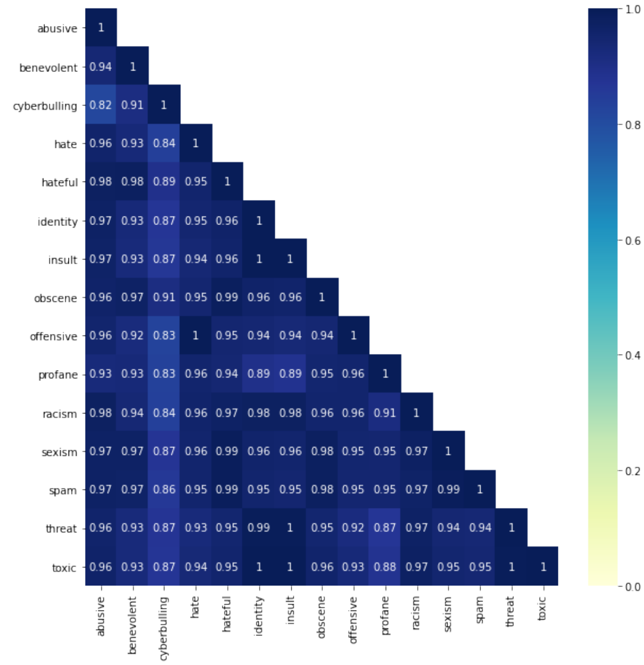

Figure 8. Cosine similarities between average representations of label embeddings.

References

[1] *hate speech*. Lexico.com.

[2] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065, 1997.

[3] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.

[4] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.

[5] Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE, 2018.

[6] Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*, 2019.

[7] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*, 2019.

[8] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social me-

dia. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666, 2012.

[9] Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain. *Social Sciences*, 9(11):188, 2020.

[10] Anna Schmidt and Michael Wiegand. 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, 2017.

[11] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16, 2017.

[12] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[13] Uwe Bretschneider and Ralf Peters. Detecting cyberbullying in online communities. 2016.

[14] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

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

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[18] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.