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 21 different subgroups of offensive language. 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 from 7 different data sets. First, we use classic approaches, such as TF–IDF, BoW, and LDA, then we move on to more sophisticated embeddings. We use both, non-contextual embeddings, such as Word2Vec, GloVe, and FastText, and contextual embeddings, such as BERT, KeyBERT, and USE. We find out that all subgroups are tightly connected in the sense of contextual embeddings. Considering all results, we provide a possible ontology of offensive language.

Keywords

Offensive language, hate speech, natural language processing, traditional methods, contextual and non-contextual embeddings, exploratory analysis, ...

1. Introduction

In the last few years social media grew exponentially and with it also the ability of people to express themselves online. Enabling people to write on different online platforms without even identifying themselves lead to a new era of freedom of speech. Despite this new medium for communication bringing 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 offensive language, 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 definition 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 21 subgroups of offensive language - *abusive, hateful, spam, general hate speech, profane, offensive, cyberbullying, racism, sexism, vulgar, homophobic, slur, harassment, obscene, threat, discredit, insult, hostile, toxic, identity hate* and *benevolent sexism*. The goal of this paper is to explore offensive language subgroups and understand the similarities and connections between them.

There has been done a lot of research regarding the offen-sive language, however, these works are usually focused on the classification. One of the first works include [2] who built the decision tree based classifier Smokey for abusive message recognition 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] use convolutional neural networks for classification into four predefined categories, and [5] use different natural language processing techniques for expanding data sets 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 also a lot of related work that focuses on creating large data sets, such as [7], who create a large-scale, multilingual, expert based data set of hate speech.

What is less common in the research area of offensive language is analysis of relationships between different types of offensive language 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 hate speech, 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.

We organize this paper as follows: we present the data sets and describe data preprocessing in Section 2, we perform the exploratory analysis by using many traditional and neural approaches in Section 3, and provide a possible offensive language ontology in Section 4.

2. Data

2.1 Data sets and label distribution

We use 7 publicly available data sets for our exploratory analysis. We combine data sets [3], [11], and [12] into one large data set (referred to as SRB) as they include same categories of hate speech. We make labels *sexism*, *racism*, and *both* from [3] and [11]. The third data set [12] that we use contains label *hostile sexism*, where marked tweets are already included in the first two data sets under *sexism*, and label *benevolent sexism*, which we rename to *benevolent*. We obtain a data set with 6069 samples that are labeled either *sexism*, *racism*, *both*, or *benevolent*. The fourth data set (referred to as AHS)[13] that we use has 3 categories - *abusive, hateful, spam*. As this is the original data set, no additional merging is needed. We obtain a data set with 13776 tweets with the mentioned labels. Note that we exclude *None* label from both data sets, as we do not need it for the analysis. We show the distribution of individual categories from data sets 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. #fuck-thanksgiving

> *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 data set of comments extracted from the League of Legends community [14]. We preprocess the data set 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. The sixth data set
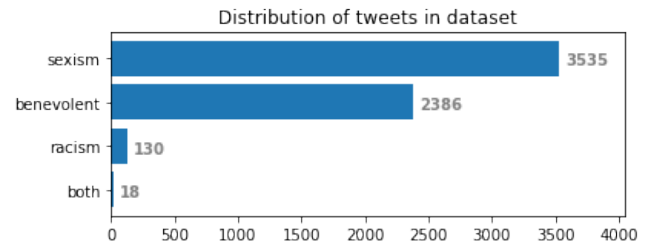

Figure 1. Distribution of tweets in SRB data set. This figure shows the distribution of hate speech categories in the SRB data set. We can see that *sexism* and *benevolent* are well represented, whereas *racism* and *both* are far less frequent. Original set contains more tweets labeled *racism*, but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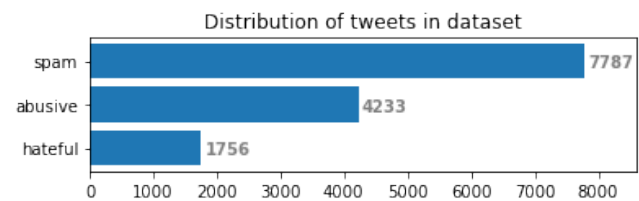

Figure 2. Distribution of tweets in AHS data set. We see that the *spam* is the most represented label in the data set, which represents the majority of the data set. This is followed by the *abusive* tweets and there is the least of *hateful* tweets. We can see that categories in this data set are well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

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 data set that we use for our analysis. It provides tags of *hatespeech, profane*, and *offensive*, so we refer to the data set as HPO. It 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

We also use the data set of Wikipedia comments [15] that are marked as either *toxic*, *severe toxic*, *obscene*, *identity hate*, *threat*, and *insult*. We merge the first two categories into *toxic*. It is important to note that each comment in this data set might have multiple labels, so the results for those tags might be similar. Original data set contains 159571 tweets, 16225 of which are labeled. We show the distribution of the labels in Figure 4. We denote this data set as TOITI in the future text.
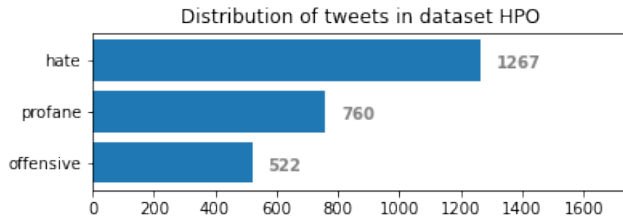
Figure 3. Distribution of tweets in HPO data set. The most used label is *hatespeech*. It is followed by *profane* and then *offensive*, which have a similar number of tweets.
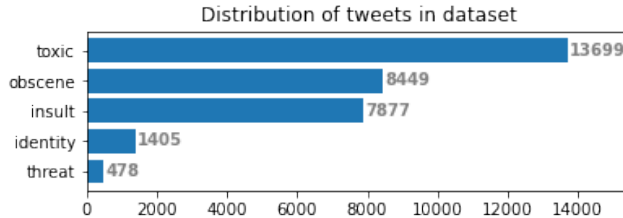

Figure 4. Distribution of tweets in TOITI. We see that most of the comments are labeled as *toxic*. Around half of them are *obscene* and around half are also labeled as *insult*. *Identity hate* and *threat* are far less common in this data set.

> *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.

2.2 Annotation process of the data

As the goal of this report is to inspect deeper structure and gain a new understanding of relationships between different subgroups of hate speech, we must also inspect how the data that we work with, was annotated. Annotations play a big role in this analysis, as we take them as a ground truth, meaning if in the data set some tweet / comment was labeled as e.g., *sexism* we do not further question this choice, and perform all our further analysis accordingly.

Data set [3] uses both amateur annotators from crowdsourcing platform CrowdFlower and annotators with theoretical and applied knowledge of hate speech, and use the data set for hate speech detection and classification. [12] manually annotate their data set with the help of a 25 year old woman studying gender studies and use the data to investigate how different is benevolent sexism from sexism, and also perform classification with SVM. [13] again use amateur annotators from CrowdFlower and want to provide large annotated data

set that is available for further scientific exploration. [14] use 3 human experts for the annotation and then propose an approach to precisely detect cyberbullies and also provide metrics to identify victims of severe cyberbullying cases. [16] used junior experts for language and they engaged with an online system to judge the tweets. Their goal was text classification. [15] again use platform CrowdFlower, however, they require their annotators to first pass a test of 10 questions to ensure data quality. Their goal is to provide methodology that will allow them explore some of the open questions about the nature of online personal attacks.

2.3 Preprocessing

Before applying any methods we first preprocess all of our data. We remove retweet text RT, hyperlinks, hashtags, taggings, new lines, and zero length tweets. We further filter out tokens that do not contain letters, e.g., raw punctuation.

3. Methodology

3.1 Traditional approaches

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

3.1.1 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 data sets. 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*.

3.1.2 TF-IDF

We continue with the analysis of data sets 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 data set. 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 data sets, as well as people not clearly differentiating between these categories. Most data sets 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 data set, 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.

3.2 Non-contextual word embeddings

For each of the category labels we try to find the 30 most similar words and use their embeddings to infer the similarities and differences between the subgroups. For this task we use pre-fitted Word2Vec [17, 18], GloVe [19], and FastText [20]. 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*, *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. 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
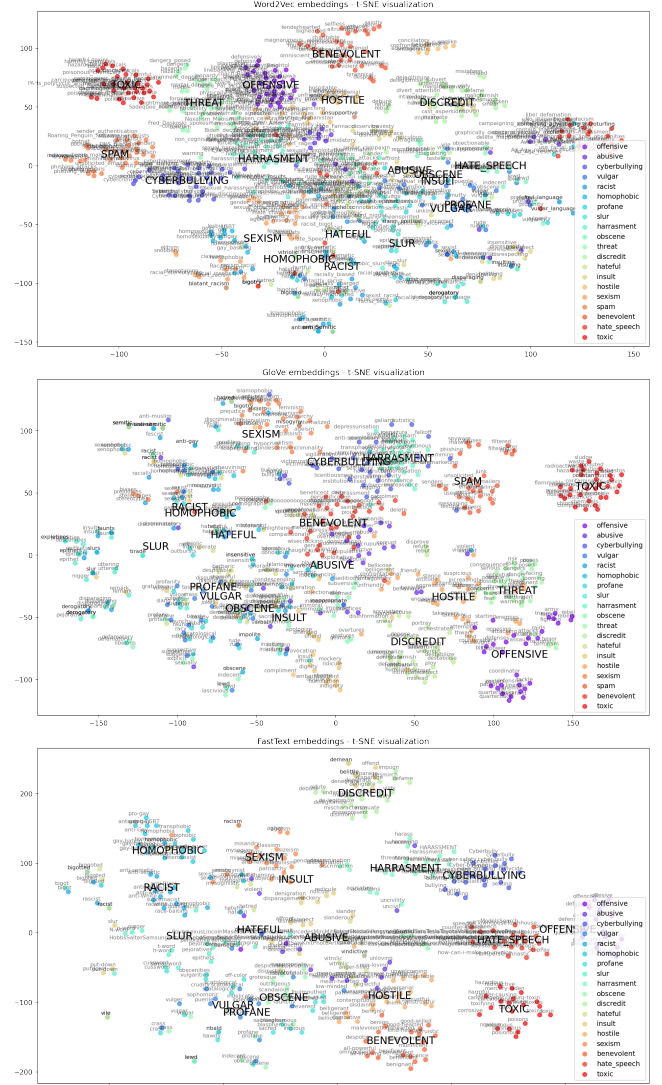

Figure 5. Word2Vec, GloVe, and FastText embeddings. Figure shows Word2Vec (1st row), GloVe (2nd row), and FastText (3rd row) embeddings of 30 closest words of each label that we analyze. Note that we omit offensive language subgroups that are not in the vocabulary.

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 pre-

cise 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, 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 offensive language 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 want to form more diverse and meaningful clusters than just 2 big subgroups as the silhouette score suggests. See the example output of the silhouette score in Figure 6.
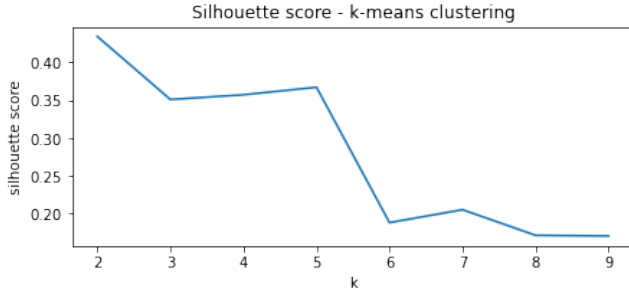

Figure 6. Silhouette score. Example of silhouette scores for different numbers of clusters. We use the second peak ($k = 5$) instead of first ($k = 2$), as we want to get more clusters.

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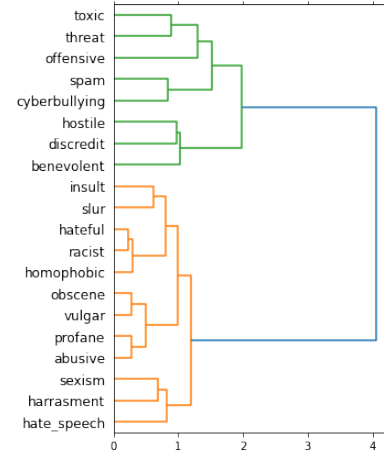

Figure 7. Hierarchical clustering of average Word2Vec embeddings of labels' 30 nearest words. Figure shows results of hierarchical clustering of the labels from data sets. Distance between two labels is computed as $1 - simcos$, where *simcos* is a cosine similarity between two labels. Embedding for each label is computed as an average of embeddings of label's nearest 30 words.

Looking at k-means clustering of Word2Vec and GloVe embeddings we see that labels *abusive*, *vulgar*, *racist*, *homophobic*, *profane*, *slur*, *obscene*, *hateful*, *insult*, and *discredit*, *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, harassment, 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.

3.3 Contextual word embeddings
3.3.1 BERT
We move on to contextual embeddings and we focus on BERT. We use the pretrained BERT base cased model [21] and convert tweets and comments from our data set to BERT embeddings. We first append them " `- This is <label>`" and compute the embeddings. From the obtained embed-
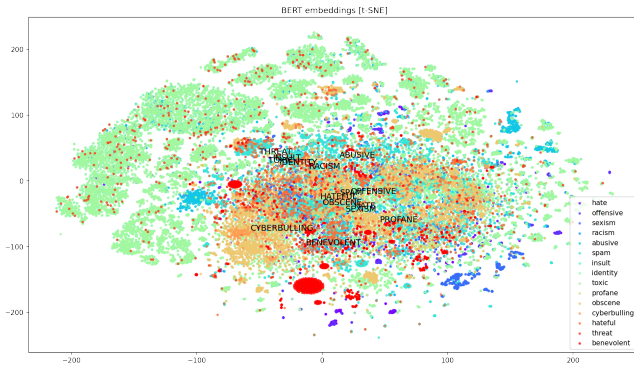
Figure 8. BERT embeddings. T-SNE visualization of BERT embeddings for different labels. We obtain each embedding by first appending `– This is <label>` to our tweets / comments and computing the embeddings for each text. An embedding of the label of text is then the average of the token embeddings that belong to the `<label>`.

Figure 9. Similarities between BERT embeddings. Figure shows similarity between labels' BERT embeddings. For each label we obtain an average vector representation by averaging embeddings obtained from label's tweets / comments (same as in Fig. 8). Similarity is then computed as a cosine similarity between those vector representations.

dings 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 9. 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. We also visualize the embeddings with the help of t-SNE in Figure 8 and we show the labels on the mean points of each subgroup. We can see that all subgroups are tightly connected and it is hard to distinguish between them. However, we can see that *cyberbullying* is a little bit more compact and not as dispersed as others, which might be a reason behind slightly different similarity scores. It is also interesting that some labels, although being dispersed, have some small clusters which stand out and might indicate special subgroups within those subgroups of hate speech. Example of such subgroup is *benevolent sexism*.

3.3.2 KeyBERT

We leverage the KeyBERT [22], which is a minimal keywords extraction technique that uses BERT embeddings to create keywords and keyphrases that are most similar to a document. For each label we compute top 3 keywords for each tweet / comment using KeyBERT, and show the labels' 5 most common keywords in Table 3. We can see that *insult*, *obscene*, and *toxic* have the same 5 most common keywords. Since they come from the same data set, and since each tweet from that data set could have multiple labels, we feel that this affected the results. We can see that quite a few labels include common keywords such as fuck, bitch, fucking, and idiot, which is not surprising, as they are among the top common curses. We can see more Trump related words in *offensive*, *profane*, and
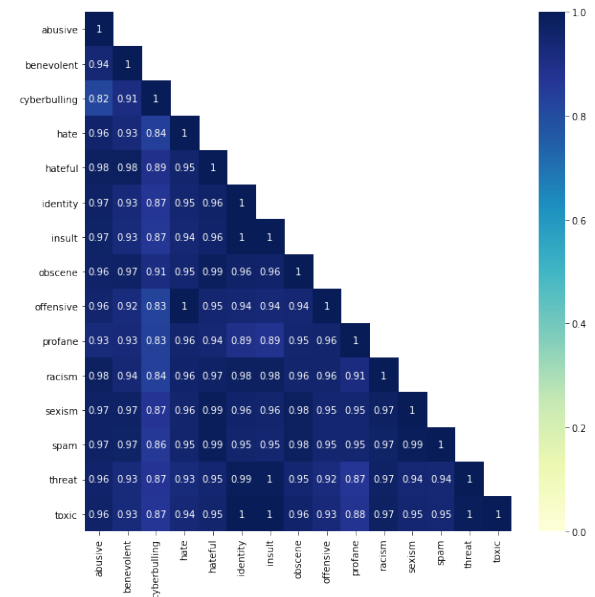
hate speech, which is probably again due to the background of data set generation. However, the most common keyword sets of those labels still slightly differ. Keywords are the most diverse between *benevolent*, *cyberbullying*, *racism*, *sexism*, and *spam*.

3.3.3 Universal Sentence Encoder

We additionally use one approach that was not suggested and that is Universal Sentence Encoder (USE) which is a model that can be nicely used for semantic similarity. As we are interested in the relations and connections between subgroups of offensive language, we try to use this model to further analyze the structure of offensive language in general.

We average the obtained embeddings of texts for each label and use cosine similarity to compute the similarity between those label representations. We show the obtained similarity matrix in Figure 10.

From the plot we can see that similarly to BERT results, the subgroups here are again very similar. We can see that *toxic, hateful,* and *spam* are more similar to each other than to other labels. However, the similarities are still to high that would allow us to infer some meaningful relation between subgroups.

4. Schema

Considering all the results and findings from above, we can now provide the following inference. Note that all categories are tightly connected in contextual embeddings, which should be kept in mind. However, we want to provide some sort of

category	BERT keywords
racism	coon, white, black, terror, fuck
sexism	sexist, women, feminazi, girls, kat
benevolent	women, womensday, sassy, adaywithoutwomen, woman
abusive	fucking, idiot, bitch, hate, fuck
hateful	hate, trump, idiot, nigga, fucking
spam	video, new, 2017, liked, free
cyberbullying	riot, troll, hacking, trolls, hacker
hate speech	trumpisatraitor, doctorsfightback, shameonicc, borisjohnsonshouldnotbepm, trump
identity hate	gay, fuck, nigger, bitch, fucking
insult	fuck, wikipedia, bitch, fucking, suck
obscene	fuck, wikipedia, bitch, fucking, suck
offensive	trumpisatraitor, fucktrump, trump, murderer, rapist
profane	fucktrump, fuck, dickhead, trump, douchebag
threat	kill, die, fuck, bitch, rape, death
toxic	fuck, wikipedia, bitch, fucking, suck

Table 3. KeyBERT keywords. Table shows 5 most common keywords found with KeyBERT obtained from tweets / comments for each offensive language subgroup.

separation where possible, so we consider more the results that separated our subcategories of offensive language. We can define the following "blocks":

1. *sexism*, *racism*, *homophobic*, and *slur*;
2. *obscene*, *insult*, *profane*, *abusive*, *vulgar*;
3. *discredit, offensive, hostile, threat, benevolent*;
4. *toxic*;
5. *spam*;
6. *cyberbullying*.

The above list includes 16 out of 21 subgroups that we analyze. The reason behind is that *hate speech* is a subgroup that can be connected with blocks 1 and 2. Similar holds for *hateful*, which can be connected to the first three blocks, however, it is never connected to *abusive*. Another subgroup that is sometimes connected to block 1 is *harassment*.

As we see that some subgroups cannot be separated just yet (block 1, 2, and 3), we apply further analysis with Word2Vec and GloVe. We focus on the labels and use the embeddings of their 50 most similar words. For the three of the following figures we use PCA visualization, so that we can also see the distance between subgroups. In the first plot of Figure 11 we see that *racism*, *slur*, and *homophobic* are more related to each other than to *sexism*. In the second plot of Figure 11 we can see that all of the inspected subgroups are tightly connected and cannot be nicely separated, *insult*, however, slightly stands out. In the last plot of Figure 11 we can see that *discredit* is not as intertwined with *insult* and *obscene*, so we conclude that although it is related to them, it it still a bit less connected to them, than they are to each other.

From the above findings, we propose a schema of offensive language subcategories in Figure 12. Note that the schema is obtained with the described analysis and it is not confirmed by any linguist professional. All of the subgroups are also tightly connected, however, as the goal of our paper is try to provide some meaningful relations, we try to summarize our findings in a schema and show more connected groups together. We find 3 mains groups, that are shown in
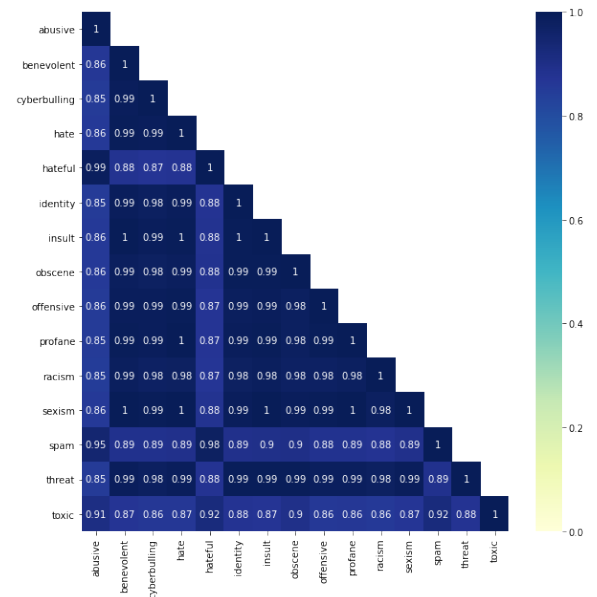

Figure 10. Similarities between USE embeddings. Figure shows similarity between labels' USE embeddings. For each label we obtain an average vector representation by averaging embeddings obtained from label's tweets / comments. Similarity is then computed as a cosine similarity between those vector representations.

bordered rectangles. Difference in colors means that the node is slightly less connected to other nodes in those groups. *Spam* and *cyberbullying* are both gray, as they are connected, but they each could be put in a separate rectangle, as they differ enough. We place them next to *toxic* as slight relations can be seen between those three. *Toxic* and *benevolent* are also connected to some of the nodes in the blue subgroup. The latter is also connected to *hateful*. We also see that *insult* has a strong relation with the red and green subgroups, and *discredit* from the blue. General *hate speech* is mostly connected to the red and green subgroup. Note that *identity hate* is related to all, but we could not find a stronger relation to any specific subgroup.

5. Discussion

Offensive language is known to everybody, as it is very common in the social media. However, we often neglect the fact that is actually a conglomerate of many subgroups, such as sexism, racism, etc. In this paper we wanted to explore offensive language and its structure and we do this by utilizing different natural language processing techniques.

We used 7 different data sets that contained Twitter and online forum comments. We used traditional techniques, such as TF–IDF, BoW, and LDA and also more advanced approaches such as non-contextual (Word2Vec, GloVe, FastText) and contextual (BERT, KeyBERT, USE) embeddings. We found out that each of the approaches provides us with slightly different relations and it is difficult to draw conclusions and we would
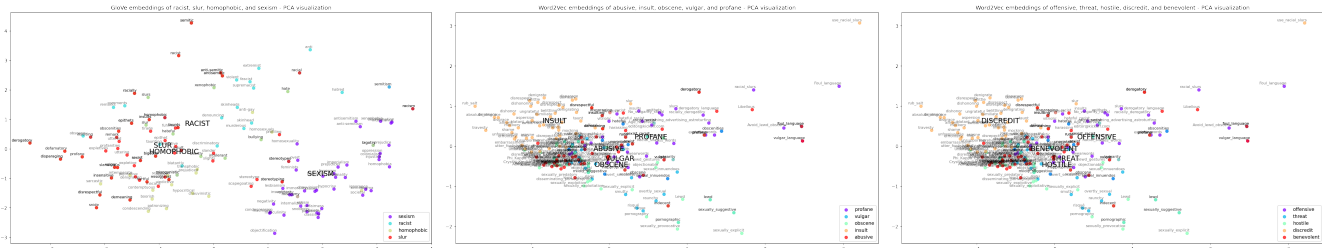
Figure 11. Further analysis of blocks that could not be separated. We use PCA visualization in all plots and 50 nearest words' embeddings of each label are used. We use GloVe embeddings in the first plot, and Word2Vec in the other two.
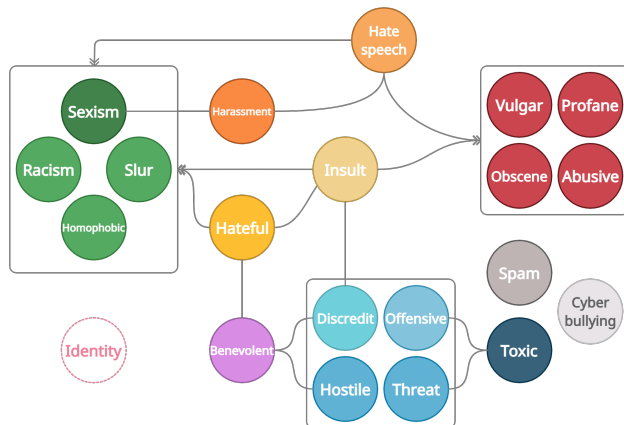

Figure 12. Proposed schema of hate speech. Figure shows the proposed schema of hate speech. Note that all of the labels are very related, however, we try to provide one possible division. Nodes in groups that are of slightly different color are more separable from other nodes in this group. We show connections to other the nodes with normal lines and connections to whole groups with line and two arrows. Identity is dotted because we do not have enough information to connect it with other node or group.

probably need some help of the linguist professionals. Results also depend on how the comments were obtained and how annotators conceive the meaning of the labels.

Combining the results from several approaches, we proposed one possible ontology of offensive language. We inferred there exist 3 groupings that include 4 subgroups of offensive language each. However, even in those groupings there exist subgroups, that are less connected to others. We also found some subgroups that are more separable from others. However, it is important to note that all the subgroups are still tighty connected.

In the future the obtained knowledge could be upgraded with the help of a linguist professional. We could also use the obtained methodology and apply it to Slovenian data sets.

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 media. 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] Zeerak Waseem and Dirk Hovy. 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, June 2016. Association for Computational Linguistics.

[12] 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.

[13] 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.

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

[15] 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.

[16] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17, 2019.

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

[18] 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.

[19] 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.

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

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

[22] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020.