Pomenska analiza kategorij sovražnega govora v obstoječih označenih korpusih

Maša Kljun, Matija Teršek, Slavko Žitnik
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, SI-1000 Ljubljana
mk2700@student.uni-lj.si, mt2421@student.uni-lj.si, slavko.zitnik@fri.uni-lj.si

Izvleček

Trenutno je dostopnih mnogo angleških korpusov z označenimi različnimi kategorijami žaljivega govora, različnimi načini označevanja in poimenovanja kategorij. V tem prispevku analiziramo 21 kategorij žaljivega oz. sovražnega govora. Pri tem uporabimo metode obdelave naravnega jezika na sedem različnih korpusih, da lahko odkrivamo korelacije med posameznimi kategorijami. Analizo izvedemo s pomočjo tradicionalnih (TF–IDF) in naprednih (fastText, GloVe, Word2Vec, BERT in ostale globoke metode) tehnik, s katerimi želimo odkriti zakonitosti med posameznimi kategorijami sovražnega govora. Rezultati razkrijejo, da je večina kategorij močno povezana med seboj, vendar lahko kljub temu izdelamo dvonivojsko hierarhično predstavitev povezanosti. Analizo izdelamo tudi za slovenski jezik in primerjamo rezultate za oba izbrana jezika.

Ključne besede: žaljivi govor, sovražni govor, obdelava naravnega jezika, vektorske vložitve besed

Abstract

There exists a vast amount of different offensive language corpora for English language, annotation criteria and category naming. In this paper, we explore 21 different categories of offensive language. We use natural language processing techniques to find correlations between the categories based on seven different data sets. We employ several traditional (TF–IDF) and advanced (fastText, GloVe, Word2Vec, BERT, and other deep NLP methods) techniques to uncover similarities among different offensive language categories. The findings reveal that most of the categories are densely interconnected, while a two-level hierarchical representation of them can be provided. We also transfer the analysis to the Slovenian language and compare the findings between both researched languages.

Keywords: Offensive language, hate speech, natural language processing, word embeddings

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* [OUP, 2021]. 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 case someone would use multiple datasets. As there exist no clear definitions of annotated categories, a researcher needs to understand them first and then decide how to use them. 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 offensive language, however, these works are usually focused on classification. One of the first works includes [Spertus, 1997] who built the decision tree based classifier Smokey for abusive message recognition and classification. Some other works that focus mainly on classification include [Waseem, 2016], who compare the classification accuracy of models trained on expert and amateur annotations, [Gambäck and Sikdar, 2017] use convolutional neural networks for classification into four predefined categories, and [Martins et al., 2018] 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 [Rizoiu et al., 2019], 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 [Chung et al., 2019], 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 the offensive language and the importance of specific keywords. Some examples include [Xu et al., 2012], who try to separate bullying from other social media posts and try to discover the topic of bullying using topic modeling with Latent Dirichlet Allocation (LDA). [Calderón et al., 2020] model hate speech against immigrants on Twitter in Spain. They try to find the underlying topic of hate speech using LDA, discovering features of different dimensions of hate speech, including foul language, humiliation, irony, etc. [Schmidt and Wiegand, 2017] conduct a survey about hate speech detection and describe key areas that have been explored, regarding the topic modeling, as well as sentiment analysis.

Recently, some research has been published, focusing on creating a new typology of offensive language [Banko et al., 2020] or trying to unify offensive language categories across datasets [Salminen et al., 2018, Risch et al., 2021]. None of these research has focused or analyzed existing data in depth. Banko et al. [Banko et al., 2020] proposed a new typology that would require re-annotation of existing data and is therefore only a theoretical ground for further annotation campaigns. Similarly, Salminen et al. [Salminen et al., 2018] propose a new taxonomy, based on existing data sources, annotate a new corpus and perform classification analysis. Risch et al. [Risch et al., 2021] try to combine a multitude of datasets into a single schema. They also provide a unification tool. We cannot agree with the analysis as we show that annotation guidelines and data sources are too much different to directly map them into one schema and that their context should be considered when doing so. We show that different categories of offensive language (as annotated in publicly available corpora) from different datasets do not have a full intersection. In the future, there is a need for comprehensive typology development, along with linguistically-sound definitions.

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. Furthermore, we use non-contextual embeddings and apply them to the Slovene language in Section 4. In the end, we provide a possible offensive language ontology in Section 5.

Note to the reader: this paper includes some explicit examples of offensive language.

2 DATA

We use 7 publicly available data sets for our exploratory analysis. We combine three data sets [Waseem, 2016], [Waseem and Hovy, 2016], and [Jha and Mamidi, 2017] into one large data set (referred to as SRB) as they include the same categories of hate speech. We create labels *sexism*, *racism*, and *both* from [Waseem, 2016] and [Waseem and Hovy, 2016]. [Jha and Mamidi, 2017] is an extension of the first two. It includes label *hostile sexism*, which contains tweets from *sexism* category in the first two data sets, and label *benevolent sexism*, which we rename to *benevolent*. Thus, we obtain a data set with 6069 samples that are labeled either *sexism*, *racism*, *both*, or *benevolent*. *Benevolent* includes comments that exhibit subjective positive sentiment, but is sexist, while *sexism* includes tweets that exhibit explicitly negative emotion. The authors do not state what was the criteria to label tweets as *racist*, but they state that it is easy to identify racist slurs.

The next data set (referred to as AHS)[Founta et al., 2018] contains 3 categories – *abusive, hateful, spam*. *Abusive* is any strongly impolite, rude, or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion. *Hateful* is language used to express hatred or is intended to be derogatory, to humiliate, or to insult the members of the group. *Spam* consists of posts related to advertising, phishing, and other kinds of unwanted information. As we use no data sets that are directly derived from this data set, contrary to the previous three data sets, we show this data set as a separate standalone data set. We obtain 13776 tweets with the above mentioned labels. Note that we exclude *None* label from both data sets, as we do not need it for the analysis. We 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 data set of comments extracted from the League of Legends community [Bretschneider and Peters, 2016], which we refer to as CYB. *Cyberbullying* is a process of sending offending messages several times to the same victim by the same offender. 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 that we use was designed for the problem of 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 [Mandl et al., 2019]. It consists of 2549 tweets. *Hateful* includes messages that describe negative attributes of individuals because they are members of a group or hateful comments towards race political opinion, gender, etc. *Offensive* includes messages that are degrading, dehumanizing, or insulting to an individual, and *profane* includes messages that contain unacceptable language in the absence of hate and offensive content (for example swearwords). We 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 [Wulczyn et al., 2017, Borkan et al., 2019] that are marked as either *toxic*, *severe toxic*, *obscene*, *identity hate*, *threat*, and *insult*. We merge the first two categories into *toxic*. Most labels here are derived from toxicity, which is defined as anything that is rude, disrecpectful, or unreasonable that would make someone want to leave a conversation. 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. The original data set contains 159571 tweets, 16225 of which are labeled. We denote this data set as TOITI in the future text and show the examples for each label:

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.

We show the distribution of individual categories from data sets in Figure 1. Note that the numbers of samples might not match the numbers in the original papers, due to the removed tweets by Twitter, making them unavailable for us to analyze. We see that *toxic, obscene, insult,* and *spam* are far more frequent than other labels, especially compared to *threat, racism*, and *cyberbullying*. This varies as the comments

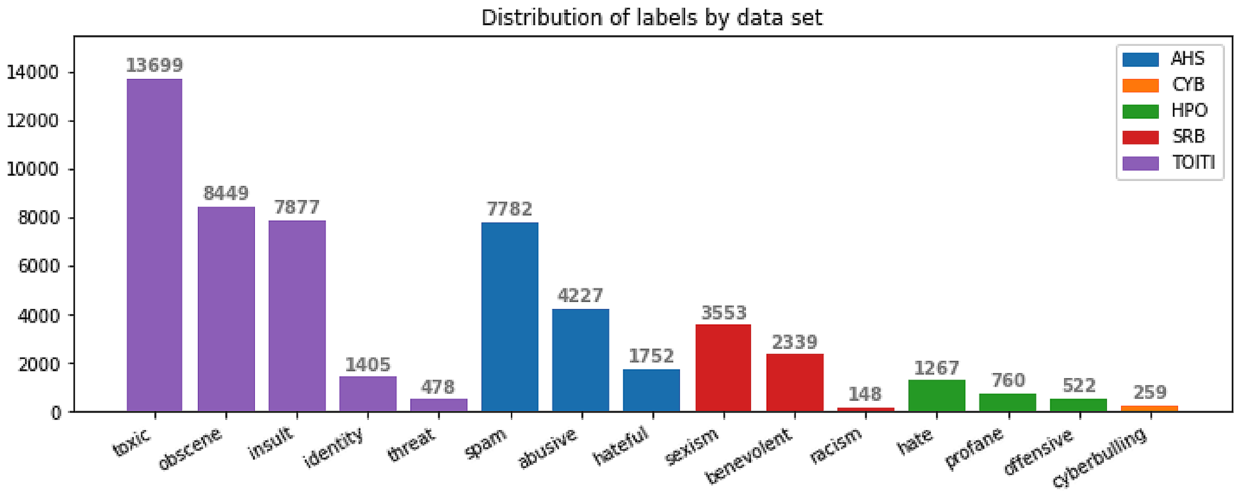

Figure 1: **Distribution of labels by data sets. We analyze the following data sets: AHS [Founta et al., 2018], CYB [Bretschneider and Peters, 2016], HPO [Mandl et al., 2019], SRB [Waseem, 2016, Waseem and Hovy, 2016, Jha and Mamidi, 2017] and TOITI [Wulczyn et al., 2017].**

were extracted from various social media platforms, which sometimes ban or remove inappropriate comments, making them unavailable for us to analyze. The number of comments for each label also depends on the size of the data set – for example, TOITI is much bigger than HPO. Note that two labels are similar (*hateful* and *hate*), and authors of both data sets use them to classify hate speech oriented towards certain groups because of their social status, disability, race, religion, ethnic origin, or sexual orientation. However, we do not merge those two labels as data sets are collected from Twitter or Facebook at different times, which might influence their content.

In addition to the 15 labels from the above mentioned data sets, we also consider six more offensive language subgroups *discredit, harassment, vulgar, homophobic, slur,* and *hostile*, which were not in the original five data sets that use. We included those words based on previous analysis done with experts from the linguistics field [Lewandowska-Tomaszczyk et al., 2021]. In this paper, we want to additionally support the claim that category naming in existing offensive datasets is not sound and therefore we cannot clearly distinguish them also using exploratory analysis tools.

As the goal of this report is to inspect the 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 were annotated. Annotations play a big role in this analysis, as we take them as ground truth, meaning if in the data set some tweet or comment was labeled as e.g., *sexism* we do not further question this choice and perform all our further analysis accordingly. The used data sets were sampled from different social mediums in a limited period at different times, and in some cases, for a specific topic (e.g., political topic). This influences the analysis. However, as the goal of this paper is to research the connections between various subgroups of hate speech, we do not question whether the data sets are a good representation of the subgroups, yet we are aware of this and keep this in mind during the analysis.

Data set [Waseem, 2016] 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. [Jha and Mamidi, 2017] 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. [Founta et al., 2018] again use amateur annotators from CrowdFlower and want to provide large annotated data set that is available for further scientific exploration. [Bretschneider and Peters, 2016] use three 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. [Mandl et al., 2019] used junior experts for language and they engaged with an online system to judge the tweets. Their goal was text classification. [Wulczyn et al., 2017] again use platform CrowdFlower, however, they require their annotators to first pass a test of ten questions to ensure data quality. Their goal is to provide a methodology that will allow them to explore some of the open questions about the nature of online personal attacks.

3 EXPLORATORY ANALYSIS

In this section, we show the analyses of the offensive language corpora. We especially focus on known NLP techniques that would help us differentiate between existing offensive language categories that are annotated in the corpora. Our analysis is conducted as follows: (A) First we employ traditional methods such TF–IDF to gather common keywords for the existing categories. (B) We continue using pre-trained and custom-trained non-contextual word embedding techniques. These enable us to gather a number of relevant vectors and then embed them into two dimensions to investigate possible differences or clusterings. (C) Lastly, we use three different contextual word embedding techniques to check for more fine-grained similarities.

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.

3.1 Traditional word embeddings

As the results using Latent Dirichlet Allocation in combination with Bag-of-Words (BoW) and TF–IDF do not add a contribution to the analysis, we employ TF–IDF as we want to see the most relevant words for each category of offensive language that we have in the data set. For each category, we take the corresponding tweets or comments and use them as documents. 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 since this can be influenced by the popular topics at that time, and a platform from which the creators of the data set decided to collect the data.

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

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

3.2 Non-contextual word embeddings

For each of the category labels, we try to find the 30 most similar words and use their embed- dings to infer the similarities and differences between the subgroups. For this task we use pre- trained Word2Vec [Mikolov et al., 2013a, Mikolov et al., 2013b], GloVe [Pennington et al., 2014], FastText [Bojanowski et al., 2017], and ConceptNet Numberbatch [Speer et al., 2017] embeddings of dimensionality 300. We visualize the results with the help of t-SNE [Van der Maaten and Hinton, 2008] (perplexity = 15, number of iterations = 3500, and 2 components). Because of this, we cannot interpret distances between the labels from the visualization. However, we can still infer that the

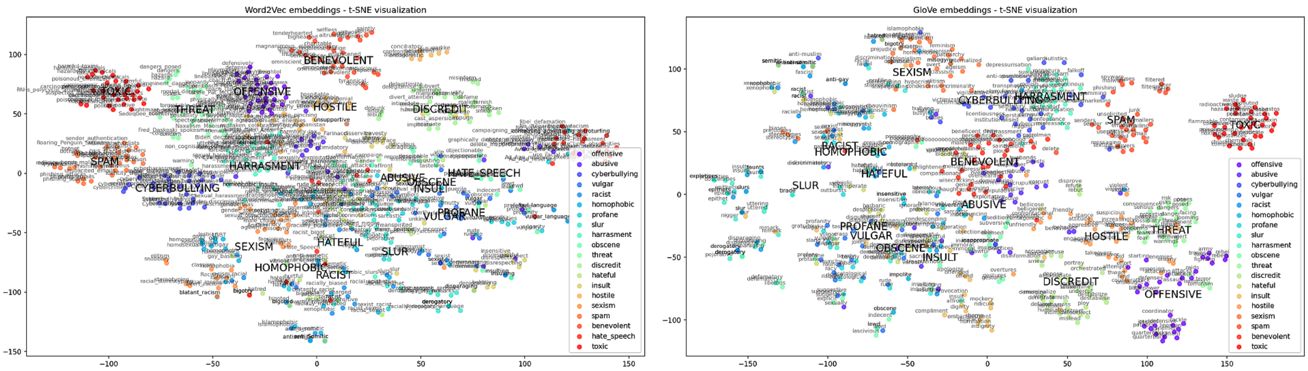

Figure 2: **Word2Vec and GloVe embeddings. Figure shows Word2Vec (left) and GloVe (right) embeddings of 30 closest words of each label that we analyze. Note that we omit offensive language subgroups that are not in the vocabulary.**

labels that are intertwined are more similar than those that are nicely separable from one another.

We show the results of Word2Vec and GloVe in Figure 2. Note that with this approach, the name of the category is favoured as the used words are derived with respect to the category name. However, the approach still uncovers various connections. 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 (*identity hate* in all models, *hate speech* in GloVe, and *threat* and *spam* in FastText).

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. We look at most similar words to the vector, which we obtain by taking the difference of unit-normed vectors of the two words on the left side of the equation and adding unit-normed vector of our_label. We consider cosine similarity.

Unfortunately, the relationships are not clear and uniquely defined. An example is *racism* is to *sexism* what is son to father with a cosine similarity of 0.646, but *sexism* is to *racism* what is son to father with a cosine similarity of 0.648. We can once again see that the two labels are related, but the precise relationship cannot be inferred. Using brother and sister the similarity 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 visualiza- tion. That is why we approach this problem with clustering. We use *k*-means (10 iterations for all experiments) and hierarchical clustering (with Ward linkage on distance matrix) 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. The silhouette score is a useful metric that can be used to validate the goodness of the clustering. It can take values from -1 (clusters assigned in the wrong way) to 1 (clusters are clearly distinguished). Silhouette score is also useful for determining the optimal number of clusters, and we use it for that purpose. 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 3.

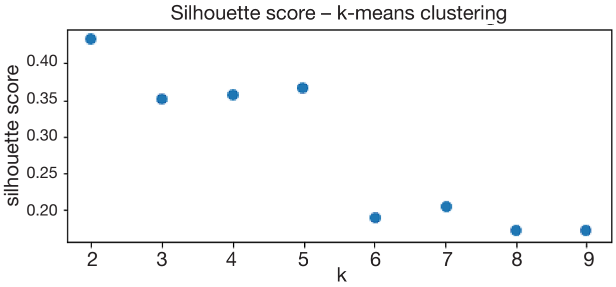

Figure 3: **Silhouette score. Example of silhouette scores for different numbers of clusters. We use the second peak (k = 5) instead of first (k = 2), as we want to get more clusters.**

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 4 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 in the same subcluster of nine offensive language groups in hierarchical clustering. They are also very similar according to the results from TF–IDF as seen before. We can see that *cyberbullying* and *spam* are clustered together in both clusterings and that *threat* and *toxic* are also very similar.

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 like those in Figure 4, so we do not show figures with those results.

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 the highest TF–IDF scores from each subgroup, however, the obtained clusters provide no useful understanding, so we omit those results.

Additionally, we use ConceptNet Numberbatch [Speer et al., 2017] embeddings. ConceptNet Numberbatch is a snapshot of word embeddings that have semi-structured, common sense knowledge from ConceptNet, a freely available semantic network. We apply a similar methodology as for Word2Vec, GloVe, and FastText embeddings, and show the results in Figure 5 using t-SNE. We can see that some subgroups are separable from the others, such as *benevolent*, *hostile*, *threat*, *homophobic*, and *spam*. We can also separate a cluster of *vulgar, obscene,* and *profane*. Other subgroups of offensive language are mainly intertwined and inseparable.

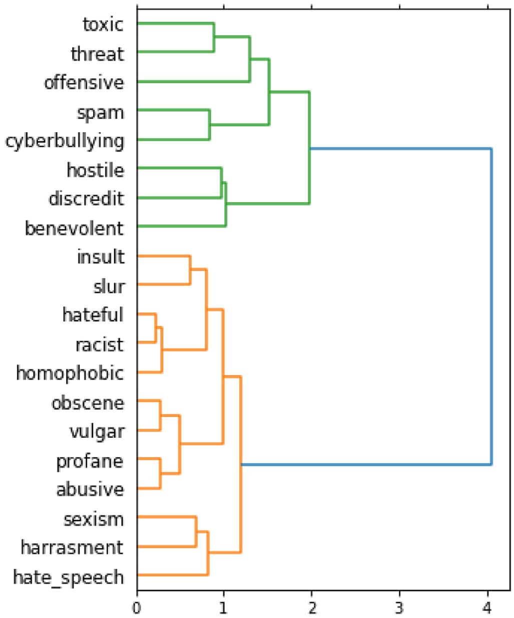

Figure 4: **Hierarchical clustering of average Word2Vec embeddings of labels' 30 nearest words. Fig- ure shows results of hierarchical clustering of the labels from data sets. Distance between two labels is computed as 1 simcos, where simcos is a cosine similarity between two labels. Embedding for each label is computed as an average of embeddings of the label's nearest 30 words.**

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.**

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

3.3 Contextual word embeddings

To perform analysis using contextual word embeddings, we need to provide whole utterances to get desired embedding vectors. We evaluate three different approaches based on BERT (Section 3.3.1, KeyBERT (Section 3.3.2) and USE (Section 3.3.3). For the plain BERT language model, we attach a category keyword to an utterance to get its representation. KeyBERT allows for automatic extraction of keywords from utterances and these represent each category. For the USE we compute average vectors from utterances and compare similarities between categories (such approach was not successful with BERT).

3.3.1 BERT

We move on to contextual embeddings and we focus on BERT. We use the pretrained BERT base cased model [Devlin et al., 2019] with 768 dimensional embeddings, 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 last-layer 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 6. 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 7 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. An example of such a subgroup is *benevolent sexism*.

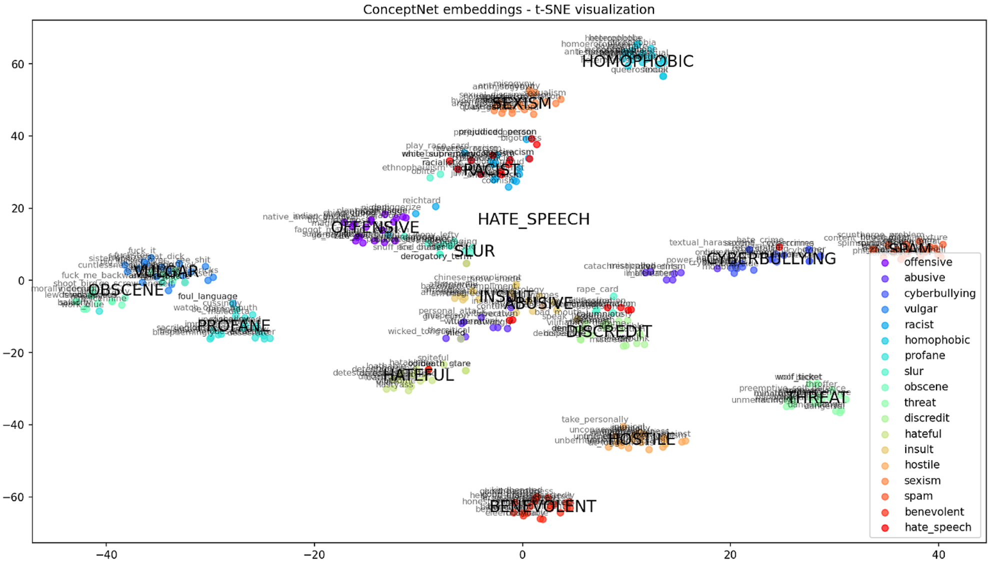

Figure 5: **ConceptNet Numberbatch embeddings. Figure shows ConceptNet Numberbatch embeddings of 30 closest words of each label that we analyze. Note that we omit offensive language subgroups that are not in the vocabulary.**

3.3.2 KeyBERT

We leverage the KeyBERT [Grootendorst, 2020], which is a minimal keywords extraction technique that uses BERT embeddings to create keywords and key phrases that are most similar to a document. For each label, we compute top three keywords for each tweet or comment using KeyBERT, and show the labels' five most common keywords in Table 3. We can see that *insult*, *obscene*, and *toxic* have the same five 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 (USE)

Another model that we use is Universal Sentence Encoder (USE) [Cer et al., 2018] which is a model that can be nicely used for semantic similarity. USE encodes text into high dimensional vectors that can be used for text classification, semantic similarity, clustering, and other natural language tasks. USE can be trained using Transformer encoder architecture [Vaswani et al., 2017] with Deep Averaging Network (DAN) [Iyyer et al., 2015]. Both models focus on a trade-off between accuracy and computational resource requirement. While the one with Transformer encoder has higher accuracy, it is computationally more intensive. For this analysis, we use universal-sentence-encoder-large model available from TensorFlow Hub, which was trained using Transformer encoder and has 512 dimensional embeddings.

We use USE 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 si-

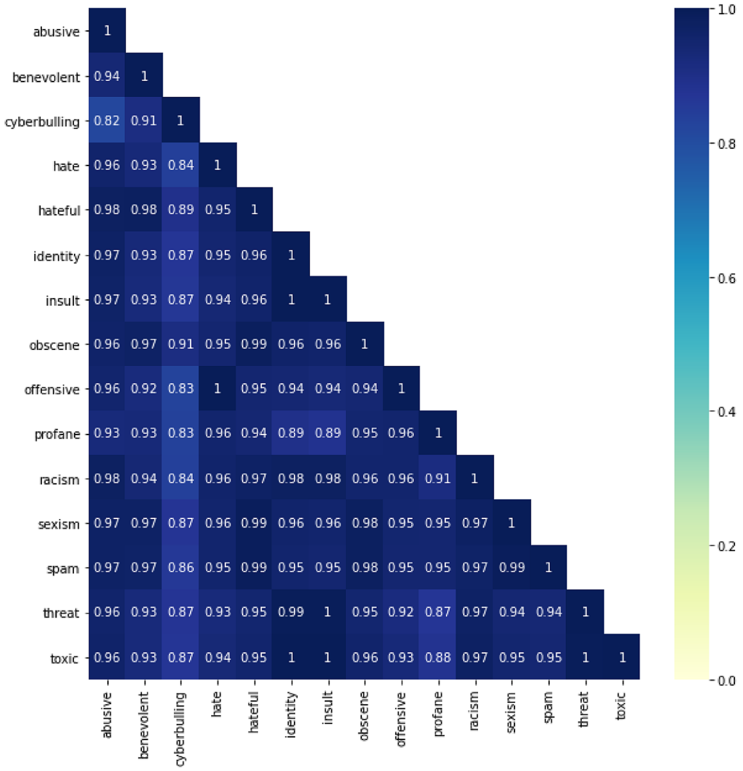

Figure 6: **Similarities between BERT embeddings. Figure shows the similarity between labels' BERT embeddings. For each label, we obtain an average vector representation by averaging embeddings obtained from the label's tweets or comments (same as in Fig. 7). The similarity is then computed as cosine similarity between those vector representations.**

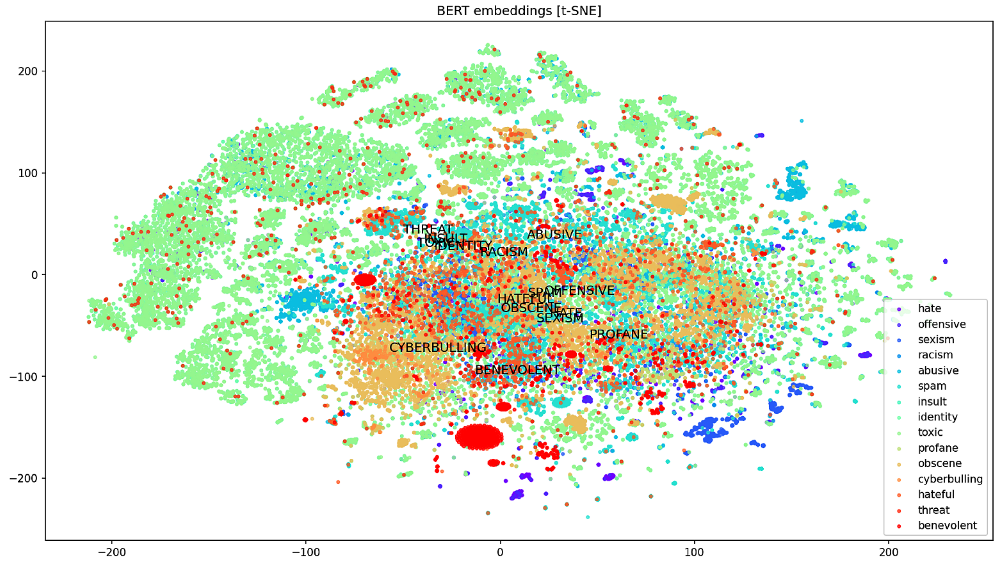

Figure 7: **BERT embeddings. T-SNE visualization of BERT embeddings for different labels. We obtain each embedding by first appending – This is <label> to our tweets or comments and computing the embeddings for each text. An embedding of the label of text is then the average of the token embeddings that belong to the <label>.**

milarity matrix in Figure 8. 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.

4 OFFENSIVE LANGUAGE IN SLOVENIAN

In this section, we translate English terms to Slovene and check whether we might uncover some differences between them using pre-trained models.

We choose to use non-contextual word embeddings. We do not focus on contextual word embeddings, as no Slovene data sets that would cover most of our labels exist. We use pretrained Word2Vec [Kutuzov et al., 2017] and FastText [Grave et al., 2018] models for Slovene language and want to see whether we can separate some subgroups of hate speech or find some subgroups that are inseparable. We first translate the labels of subgroups into Slovene language and we show the translations in Table 4. We intentionally translate all labels to nouns in order to keep them all in the same part of speech, as experiments showed that otherwise the labels that shared the same part of speech were intertwined. Unfortunately, as some words are not supported in Slovenian Word2Vec and FastText, we remove labels for hate speech (*slo. sovražni govor*), spam (*slo. vsiljenost*), and cyberbullying (*slo. spletno nasilje*) for Word2Vec and hate speech (*slo. sovražni govor*), toxic (*slo. toksičen*), and cyberbullying (*slo. spletno nadlegovanje*) for FastText. Although FastText supports word-parts, the splits did not include meaningful roots of the keywords and therefore we ommit them from results. For each of the supported category labels we try to find the ten and twenty most similar words for

Table 3: **KeyBERT keywords. Table shows five most common keywords found with KeyBERT obtained from tweets or comments for each offensive language subgroup.**

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

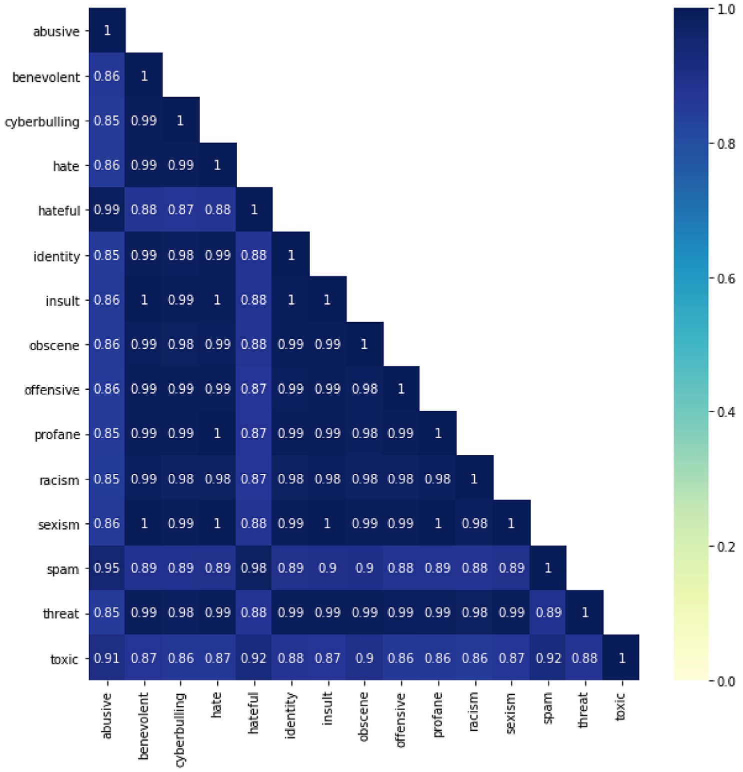

Figure 8: **Similarities between USE embeddings. Figure shows the similarity between labels' USE embeddings. For each label, we obtain an average vector representation by averaging embeddings obtained from the label's tweets or comments. The similarity is then computed as cosine similarity between those vector representations.**

Word2Vec and FastText models, respectively, and use their embeddings to infer the similarities and differences between the subgroups. We show the results of Word2Vec and FastText in Figure 9. We can see from Word2Vec visualization that toxic (*slo. toksičen*) is the only subgroup that can be well separated from others while all other subgroups are inseparable. Inspecting the FastText t-SNE visualization, we see that the only well separable subgroup is homophobic (*slo. homofobija*). Otherwise, there exist three groups that contain two or more subgroups of offensive language that are inseparable. An example of such a group is one smaller group that contains racism (*slo. rasizem*) and sexism (*slo. seksizem*) while two other groups contain five and eight subgroups of offensive language, respectively.

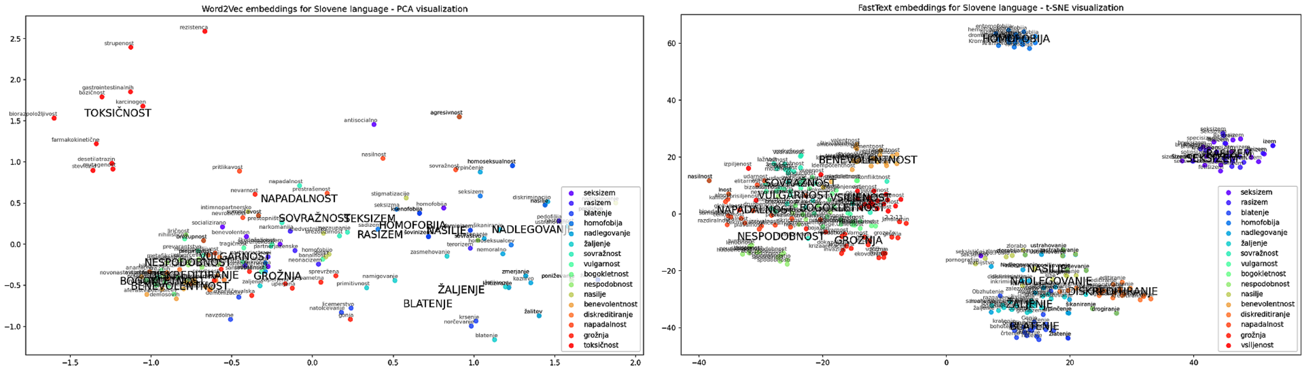

Figure 9: **Analysis of offensive language ontology for the Slovene language. Figure shows PCA visualization of Word2Vec and t-SNE visualization of FastText embeddings.**

5 DISCUSSION

Considering all the results and findings from above, we can now provide the following inference. Note that all categories are tightly connected in the results of contextual embeddings, which should be kept in mind. However, we want to provide some sort of separation where possible, so we consider more those results that separated our subcategories of offensive language. From all of the performed analysis, we can conclude that *spam* and *cyberbullying* can both be separated from other subcategories. We put *toxic* as a separate block as it is distinguishable from others in Word2Vec and GloVe embeddings, however, from clustering results we can see that it can also be connected to *offensive* and *threat*. We put *obscene*, *insult*, *profane*, *abusive*, and *vulgar* together as they appear in the same cluster in k-means clustering of Word2Vec and GloVe embeddings, and as they have quite similar words in KeyBERT. We define the remaining two subgroups by inspecting the Word2Vec results. Thus we obtain 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*.

Table 4: **Slovenian translation of labels. Table shows English labels and their Slovenian translations. We only show labels for which we found a suitable translation. We use only benevolenten as a translation for benevolent sexism, as it is mostly used in connection with benevolenten seksizem.**

English word	Slovene translation
Sexism	Seksizem
Racism	Rasizem
Slur	Blatenje
Homophobic	Homofobija
Hate speech	Sovražni govor
Harassment	Nadlegovanje
Insult	Žaljenje
Hateful, hostile	Sovražnost
Vulgar	Vulgarnost
Profane	Bogokletnost
Obscene	Nespodobnost
Abusive	Nasilje
Benevolent sexism	Benevolentnost
Discredit	Diskreditiranje
Offensive	Napadalnost
Threat	Grožnja
Toxic	Toksičenost
Spam	Vsiljenost
Cyberbullying	Spletno nadlegovanje

In the above list, we only state 17 out of 21 subgroups that we analyze, as some categories could be tightly connected to multiple subgroups. As 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. We use PCA visualization (with 2 components), so that we can also see the distance between subgroups. In the first plot of Figure 10 we see that *racism*, *slur*, and *homophobic* are more related to each other than to *sexism*. In the second plot of Figure 10 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 10 we can see that *discredit* is not as intertwined with *insult* and *obscene*, so we conclude that although it is related to them, it is less they are to each other.

From the above findings, we show a schema of offensive language subcategories in Figure 11. 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 to provide some meaningful relations and ontology, we try to summarize our findings in a schema and show more connected groups together. We find 3 main groups, that are shown in bordered rectangles. The 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 relationship with the red and green subgroups, and *discredit* from the blue group. 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.

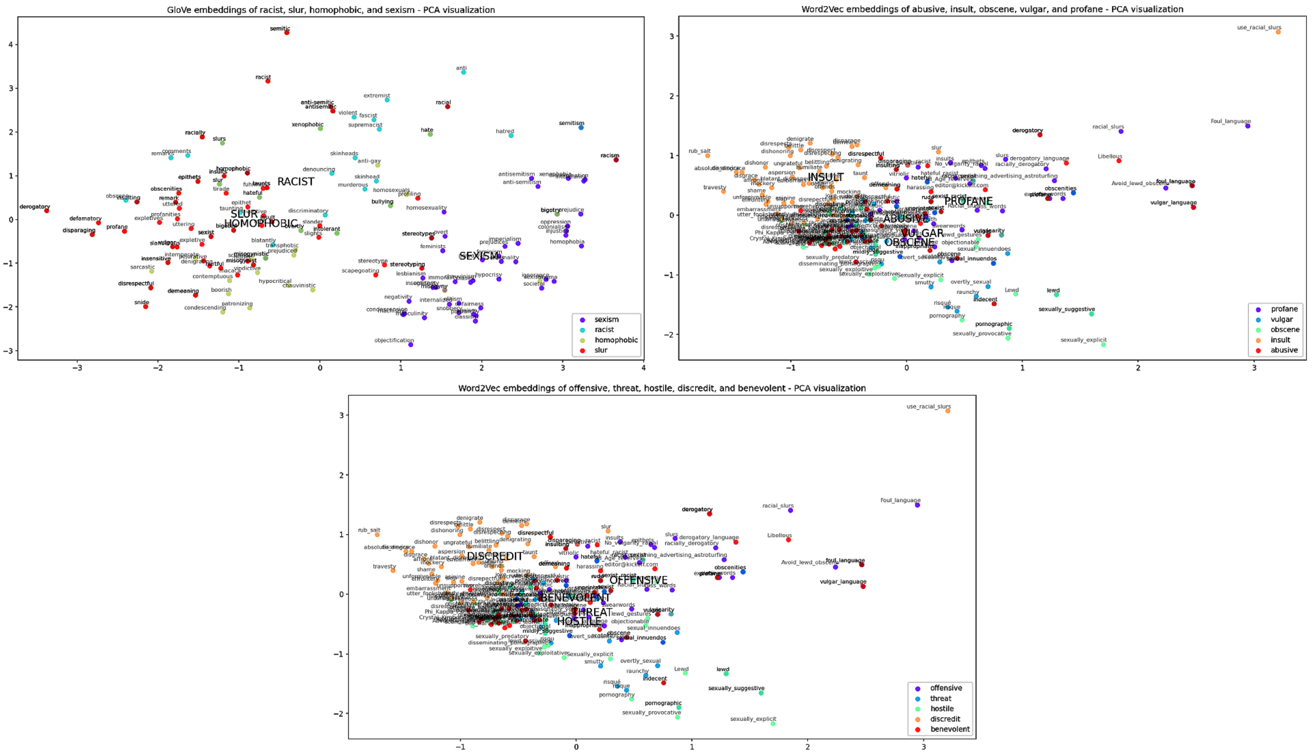

Figure 10: **Further analysis of blocks that could not be separated. We use PCA visualization in all plots and 50 nearest words' embeddings of each label are used. We use GloVe embeddings in the first plot, and Word2Vec in the other two.**

We compare our taxonomy to the taxonomy defined in [Banko et al., 2020]. This is a challenging task, as the proposed taxonomy in [Banko et al., 2020] is only a theoretical ground for further annotation campaigns and not derived from data, thus containing different and missing some of the categories in our paper. The authors propose four main subcategories of online harm: Hate and harassment, self-inflicted harm, ideological harm, and exploitation. We can see that the green nodes from our taxonomy in Figure 11

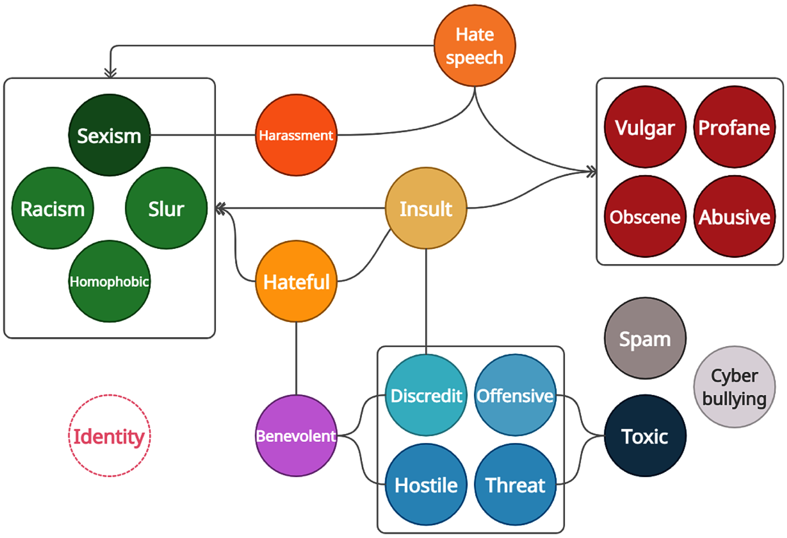

Figure 11: **Inferred schema of hate speech. Figure shows the inferred schema of hate speech. Note that all of the labels are very related, however, we try to provide one possible division. Nodes in groups that are of slightly different color are more separable from other nodes in this group. We show connections to other nodes with normal lines and connections to whole groups with lines and two arrows. Identity is dotted because we do not have enough information to connect it with other nodes or groups.**

can be classified into multiple subgroups of hate and harassment. *Spam* could be included in misinformation, which is a subgroup of ideological harm, as well as into some subcategories of exploitation. Some of the categories, like *threat, offensive,* and *profane* could be categorized into multiple subgroups of hate and harassment as well, while for some we could not find appropriate subcategories of online harm. We did not cover groups of hate speech that could be categorized into self-inflicted harm from [Banko et al., 2020], including self-harm and eating disorder promotions, or specific categories that could be categorized into subcategories of exploitation, such as child sexual abuse material or adult sexual services. Existing data is missing such annotations and relabeling could be beneficial for further exploration of subcategories of online harm proposed by [Banko et al., 2020].

6 CONCLUSIONS

Offensive language is known to everybody, as it is very common in social media. However, we often neglect the fact that is 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 seven different data sets that contained Twitter and online forum comments. We used traditional techniques, such as TF–IDF, 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 probably need some help from 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 inferred one possible ontology of offensive language. We inferred there exist three groupings that include four 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 tightly connected.

Additionally, we used pre-trained Slovenian Word2Vec and FastText models and found out that toxic (*slo. tok- sičnost*) and homophobic (*slo. homofobija)* can be nicely separable by Word2Vec and FastText, respectively. Having a Slovene data set that would cover most of our labels would also be beneficial, as we could also use contextual embeddings. This would help us infer an ontology and we, therefore, delegate this to future work.

In the future, the obtained knowledge could also be upgraded with the help of a linguist professional. Only a few data sets for Slovene offensive language exist at the moment. These include Slovenian Twit- ter dataset 2018-2020 1.0 [Evkoski et al., 2021] and Slovenian Twitter hate speech dataset IMSyPP--sl [Kralj Novak et al., 2021], with labels acceptable, inappropriate, offensive, and violent, with the latter data set also containing some information to whom the hate speech is directed (LGBT, racism, sexism, homophobia, etc.), and Offensive language dataset of Croatian, English and Slovenian comments FRENK

1.1 [Ljubešić et al., 2021], which contains six categories – violence, offensive speech, threat, inappropriate speech, and acceptable speech. For future work, we see additional value in expanding those or creating new data sets, that would cover all categories analyzed in this paper. Note that we used only pretrained embeddings which were in our case too general and resulted in inseparable categories. Better results might be obtained by using more problem specific embeddings, such as HateBERT [Caselli et al., 2020].

7 ACKNOWLEDGEMENTS

We thank Žiga Trojer, Dimitar Stefanov, and Tomaž Martinčič for suggesting and help with ConceptNet Numberbatch.

REFERENCES

[1] [Banko et al., 2020] Banko, M., MacKeen, B., and Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

[2] [Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[3] [Borkan et al., 2019] Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

[4] [Bretschneider and Peters, 2016] Bretschneider, U. and Peters, R. (2016). Detecting cyberbullying in online communities.

[5] [Calderón et al., 2020] Calderón, C. A., de la Vega, G., and Herrero, D. B. (2020). 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.

[6] [Caselli et al., 2020] Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

[7] [Cer et al., 2018] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo- Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium.

[8] [Chung et al., 2019] Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). CONAN – COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819– 2829, Florence, Italy. Association for Computational Linguistics.

[9] [Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

[10] [Evkoski et al., 2021] Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., and Kralj Novak, P. (2021). Slovenian twitter dataset 2018-2020 1.0. Slovenian language resource repository CLARIN.SI.

[11] [Founta et al., 2018] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

[12] [Gambäck and Sikdar, 2017] Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

[13] [Grave et al., 2018] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[14] [Grootendorst, 2020] Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT.

[15] [Iyyer et al., 2015] Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 1681–1691.

[16] [Jha and Mamidi, 2017] Jha, A. and Mamidi, R. (2017). 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.

[17] [Kralj Novak et al., 2021] Kralj Novak, P., Mozetič, I., and Ljubešić, N. (2021). Slovenian twitter hate speech dataset IMSyPP-sl. Slovenian language resource repository CLARIN.SI.

[18] [Kutuzov et al., 2017] Kutuzov, A., Fares, M., Oepen, S., and Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 58th Conference on Simulation and Modelling,* pages 271–276. Linköping University Electronic Press.

[19] [Lewandowska-Tomaszczyk et al., 2021] Lewandowska-Tomaszczyk, B., Žitnik, S., Baczkowska, A., Liebe- sking, C., Mitrović, J., and Oleskevisiene, G. V. (2021). Lod-connected offensive language ontology and tagset enrichment. In *Proceedings of the First Workshop on Sentiment Analysis & Linguistic Linked Data*, pages 1–16.

[20] [Ljubešić et al., 2021] Ljubešić, N., Fišer, D., Erjavec, T., and Šulc, A. (2021). Offensive language dataset of croatian, english and slovenian comments FRENK 1.1. Slovenian language resource repository CLARIN.SI.

[21] [Mandl et al., 2019] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). 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.

[22] [Martins et al., 2018] Martins, R., Gomes, M., Almeida, J. J., Novais, P., and Henriques, P. (2018). Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE.

[23] [Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[24] [Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

[25] [OUP, 2021] OUP (2021). Lexico.com – Oxford University Press. https://www.lexico.com/definition/ hate_speech. Accessed: 2021-09-01.

[26] [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[27] [Risch et al., 2021] Risch, J., Schmidt, P., and Krestel, R. (2021). Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 157–163. Association for Computational Linguistics.

[28] [Rizoiu et al., 2019] Rizoiu, M.-A., Wang, T., Ferraro, G., and Suominen, H. (2019). Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*.

[29] [Salminen et al., 2018] Salminen, J., Almerekhi, H., Milenković, M., Jung, S.-g., An, J., Kwak, H., and Jansen,

[30] B. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[31] [Schmidt and Wiegand, 2017] Schmidt, A. and Wiegand, M. (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.

[32] [Speer et al., 2017] Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general kno-

wledge. In *Thirty-first AAAI conference on artificial intelligence*, pages 4444–4451.

[33] [Spertus, 1997] Spertus, E. (1997). Smokey: Automatic Recognition of Hostile Messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, page 1058–1065. AAAI Press.

[34] [Van der Maaten and Hinton, 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[35] [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[36] [Waseem, 2016] Waseem, Z. (2016). 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.

[37] [Waseem and Hovy, 2016] Waseem, Z. and Hovy, D. (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.

[38] [Wulczyn et al., 2017] Wulczyn, E., Thain, N., and Dixon, L. (2017). 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. International World Wide Web Conferences Steering Committee.

[39] [Xu et al., 2012] Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). 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.

▪

Maša Kljun is a Data Science Masters's student at the Faculty of Computer and Information Science, University of Ljubljana.

▪

Matija Teršek is a Data Science Masters's student at the Faculty of Computer and Information Science, University of Ljubljana.

▪

Slavko Žitnik is an assistant professor at the Faculty of Computer and Information Science, University of Ljubljana. His main research interests are information retrieval and information extraction. Specifically, he is trying to enrich the extracted data from text using parallel and iterative combination of entity extraction, relationship extraction and coreference resolution techniques. Furthermore, his research also focuses on data merging, redundancy elimination and ontologies.