University *of Ljubljana*
Faculty *of Computer and Information Science*

Offensive language exploratory analysis

Juš Hladnik, Miha Torkar

Abstract

In this project we focused on exploratory analysis regarding hate speech and offensive language. We used different methods for finding relations between different typed of hate speech.

Keywords

Offensive language, hate speech, exploratory data analysis

Advisors: Slavko Žitnik

Introduction

In this project we do exploratory analysis on hate speech where we explore different datasets, how they are structured and labeled and try to find importance of specific words and phrases and relationship between types of hate speech. We make an overview of similar approaches to detecting hate speech and we propose baseline methods that we will use to learn importance of specific words and relationships between them.

Datasets and literature overview

- Gomez et al. [1] create manually annotated multimodal hate speech dataset formed by 150,000 tweets, each one of them containing text and an image which they call MMHS150K. They classify each tweet in one of 6 categories: No attacks to any community, racist, sexist, homophobic, religion based attacks or attacks to other communities. They focus on detecting categories of hate speech with both text and images from tweets. First they train a LSTM on tweet text where they use pre-trained GLoVe embeddings. They improve this model by incorporating also text from images. They then use various multimodal architectures to use text and images simultaneously.
- Qian et al. [2] introduce 2 datasets with full conversations instead of single posts, which is meant to provide context. Datasets consist of 5K conversations retrieved from Reddit and 12k conversations retrieved from Gab, both of which are social media platforms. For each conversation they provide labels for posts (hate speech/not hate speech) and human written intervention responses. For binary hate speech detection they use logistic regression and support vector machines with features as TF-IDF values of up to 2-grams. They also use CNN and RNN models for sentence classification with word2vec embeddings.
- Chung et al. [3] propose a large scale multilingual dataset with hate speech type and subtype annotations for English, French and Italian. They also provide counter speech responses in an effort to counter hate speech
- de Gibert et al. [4] create a dataset with sentences from Stormfront, a white supremacist forum, and they label each sentence whether it contains hate speech or not. Dataset consists of 10568 sentences. They have then trained a classifier on this dataset to detect hate speech, where they used SVM with bag-of-words features, they also used CNN and RNN with LSTM
- Davidson et al. [5] create a dataset with 24802 posts from Twitter and they label them as 3 possible categories: hate speech, offensive language and neither of those. They use this dataset to build features such as bigram, unigram, trigram TF-IDF, and they also include features such as length of tweet, number of hashtags, mentions, retweets,... They then test a variety of models: logistic regression, naıve Bayes, decision trees, random forests, and linear, SVMs.
- Mandl et al. [6] proposed a competition in which datasets in 3 languages were provided (1 of them is English). Dataset is developed from Twitter and Facebook and every entry is binary (hateful and offensive or not) and more fine-grained (hate, offensive and profane) classified, further for each it is declared if the speech is targeted or not. English dataset consists of circa 8000 entries. For all tasks 321 experiments with different

approaches has been submitted.

- Fersini et al. [7] similarly propose a competition in which goal was to identify misogyny and a dataset was provided in English and Spanish, the English one consisting of around 4000 tweets, which were labeled as misogynistic or not. Misogynistic tweets were further classified into 5 more categories and also the target was depicted. For classification in English 11 different approaches were submitted. All used more basic models, such as SVM, Bag of Word etc.
- Wulczyn et al. [8] provide a corpus of more then 100.000 comments from Wikipedia, which are labeled by human as attacks or not and if the comment has aggressive tone. Since more people were tasked with labeling of each text, we also get a fraction of people that labeled each comment as offensive. Researchers represent data as bag-of-words and build few models (logistic regression, multi-layer perceptrons). The models are then used on more then 63 million comments so the dataset we could potentially use is huge.
- Zampierieta et al. [9] also conducted a competition in identifying and categorizing offensive language in social media and provided a dataset of 14.100 3-level annotated tweets. First level is only a label if tweet is offensive or not, second and third level are about targets of offensive speech. They experiment classification with use of SVMs, BiLSTM (bidirectional Long Short-Term-Memory model) and CNNs. For competition part[10] more then 100 different models were submitted for all 3 parts of classification problem, for first level classification most successful approaches used BERT, meanwhile on average on all 3 levels the most successful approach used ensembles.
- In blog post [11] Susan Li uses topic modelling for discovering topics and keyword extraction in unlabeled corpus. They use Latent Dirichlet Allocation for unsupervised learning of topic modelling and to get which words a re the most representative of which topic.

Finding and merging right datasets

For the assignment we have to select enough datasets to cover the required terms and in this section we describe how we selected and merged datasets together. We searched for datasets that contained records with required labels. From each dataset we extracted records with label that we needed and made a separate file with these records and at the end we merged all such documents together. We found some datasets that had labeled some records with multiple labels and therefore we can found some duplicated records with different labels in our final dataset.

- We use dataset from Kaggle[1] to get words. Dataset is composed from a large number of Wikipedia comments which have been labeled by human raters for toxic behavior (hate speech). Some comments have multiple labels. Comments are labeled as: toxic, severe toxic, obscene, threat, insult and identity hate. From this dataset we retrieve comments for required terms: obscene (8449 texts), threat (478 texts), insult (7877 texts) and additional terms: severe_toxic (1595 texts) and identity_hate (1405 texts).

- We have already described dataset from Davidson et al. [5] previously. We use it to extract labels hateful (hate speech) (1430 texts) and offensive (offensive language) (19190 texts).

- Vidgen et al. [12] present a large synthetic training dataset for online hate classification, created from scratch with trained annotators over multiple rounds of dynamic data collection. Records in dataset are labeled as hate speech or no hate speech and hate speech is divided into types: derogation, animosity, threatening. We use records with animosity label to extract texts for term hostile (2377 records) (as animosity is strong hostility by definition).

- We have already described dataset from Mandl et al. [6] previously. We use it to extract labels profane (667 texts), offensive (451 texts) and hateful (1143 texts).

- Waseem et al. [13] present a dataset with 16,914 tweets and annotations with labels: sexist, racist and none. We extract labels for terms racist (1939 texts) and sexist (3148 texts).

- Chandrasekharan et al. [14] analyzed comments deleted by Reddit moderators. They used topic modelling to map comments to the rule it was breaking and that way they prepared a document with comments that violate some rule. We use comments that contain misogynistic slurs for our key term slur (5059 texts), comments that contain abuse and criticism towards moderators for term abuse (5059 texts), and comments that contain either racism or homophobic hate speech for term homophobic (5059 texts) (this one is a bit problematic as each comment is not definitively labeled as racist or homophobic but can be or the other or both).

- Cachola et al. [15] collect a novel corpus of tweets, all of which contain vulgar words, and annotate them for sentiment. We use all of their collected tweets for term vulgar (6718 texts).

- In formspring data[2] there are posts (question plus answer) that are labeled by 3 independent labelers whether a post contains cyberbullying or not. We take all posts that were labeled as cyberbullying by all of the labelers and use it for our term cyberbullying (339 texts).

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[2] originally provided at https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection.

label	number of texts
abuse	5059
cyberbullying	339
harassment	5285
hateful	2573
homophobic	5059
hostile	2377
identity_hate	1405
insult	7877
obscene	8449
offensive	19641
profane	667
racist	1939
severe_toxic	1595
sexist	3148
slur	5059
threat	478
vulgar	6718
all	77668

Table 1. All together we have 77.668 labeled texts.

- Golbeck et al. [16] prepared a large human-labeled corpus for online harassment research where they labeled 35.000 tweets as harassment or non-harassment. We extract tweets labeled as harassment for our key term (5285 texts).

In Table 1 we can see how many of each labels we have in our final dataset, some labels occur much more often then others.

Methods

Data preprocessing

As the main sources were online social networks (such as Twitter) there was a lot of unwanted text elements in it, such as hashtags, emojis, retweet tags (RT) etc. We also had to remove some control characters as newline tags etc. and converted all text to lower case plus removed the unwanted punctuation. We mostly achieved this by using regular expressions and for removing stop words we used NLTK package.

After we cleared the texts, we tokenized them and further lemmatized and stemmed (using Snowball stemmer and WordNet lemmatizer) each token, but for later work we only used lemmatized tokens. Some examples of our final texts can be seen in Table 2.

Results

Frequency counts

Firstly we used the most primitive way and just counted how many times each word appeared in each of our labels. Results were quite as expected as in almost all labels a version of word ”fuck” is in top words. Also already here some differences can be seen as i.e. in label ”racist” there are words (”muslim”, ”mohhamed”..) that do not occur in other labels. Labels with 4 most frequent words can be seen on table 3.

If we look into if there are some words that occur in most of the labels, we quickly see, that words like ”fuck” and ”shit” occur in quite a few of them. One measure of how and if labels are similar to one another is to look into intersection between most frequent words. Such image together with some explanation can be seen on figure 1.

This way we found out, that some labels share very little with other, while on the other hand some labels have quite few most frequent words in common. Main take away here would not be the size of intersection between labels (as we did not remove word that occur in most of them), but that there are labels that do not share many common words with others, like ”racist”, ”threat” and ”sexist”.

Comparison using TF-IDF

For TF-IDF comparison we joined all texts from each label into one single document and then created TF-IDF representation of this document. Once we had that we compared vectors with use of cosine similarity, result can be seen on figure 2, here once again the relation between labels ”homophobic” and ”abuse” is visible. Another one strong connection seems to be between labels ”insult”, ”obscene” and ”severe_toxic”, all these connections are also clear when we build a dendrogram using hierarchical clustering (see figure 3). Other than clusters already seen on few more clusters are visible (”vulgar”, ”cyberbullying” and ”offensive”).

Word2vec, Glove and fastText

Pretrained Word2vec, Glove and fastText

We then used pretrained word2vec, GloVe and fastText and embedded important words for each category. We tried with embedding most frequent words and words that TF-IDF deemed most important.

We visualized each of the possible embeddings using PCA, MDS and TSNE. We also tested how number of most important words affects clustering. A few best clustering results using pretrained embeddings can be seen on Figure 4. We found that using too little words we don’t get too significant results as there is word ”fuck” as one of the most important words in most of the labels. But using too much words we would get all labels clustered together as there are is a subset of bad words that are found in all labels’ top 50 most important words.

Even though the clusters for labels are not perfectly separable we get the best results using PCA clustering and using different pretrained embeddings yield very similar results. From Figure 4 we see that there are some rough clusters formed and that the most isolated labels (meaning different than others) are ”racist”, ”sexist” and ”hostile”. On the other hand labels ”homophobic”, ”abuse” and ”hateful” seems like they are almost the same.

Word2vec and fastText trained on our corpus

We then trained Word2vec and fasttext models on our own corpus to generate more appropriate embeddings and we tried generating vector embeddings of different lengths. After that

text	tokenized_text	stems	lemmas
RT @LookinnnAss: - you a pussy I see yo whiskers!	[pussy, see, whiskers]	[pussi, see, whisker]	[pussy, see, whisker]
fucking nazi mods	[fucking, nazi, mods]	[fuck, nazi, mod]	[fucking, nazi, mod]
<USER> who in the hell cares	[hell, cares]	[hell, care]	[hell, care]

Table 2. Few examples of text preprocessing, we removed the unwanted parts of texts, tokenized, stemmed and lematized them.

label	n_most_freq_words	top_n_tfidf_words
homophobic	[fucking, fuck, child, faggot]	[fucking, fuck, trafficker, trafficking]
hateful	[bitch, faggot, like, trumpisatraitor]	[trumpisatraitor, shameonicc, faggot, doctorsfi...
slur	[cunt, fucking, fuck, idiot]	[cunt, fucking, fuck, idiot]
identity_hate	[nigger, jew, fat, gay]	[nigger, jew, fat, gay]
hostile	[people, woman, black, like]	[people, woman, black, like]
obscene	[fuck, suck, shit, fucking]	[fuck, suck, shit, fucking]
sexist	[sexist, mkr, woman, girl]	[mkr, sexist, kat, woman]
severe_toxic	[fuck, suck, as, shit]	[fuck, suck, as, shit]
insult	[fuck, suck, nigger, fucking]	[fuck, suck, nigger, faggot]
offensive	[bitch, hoe, like, pussy]	[bitch, hoe, like, pussy]
racist	[islam, muslim, mohammed, religion]	[islam, muslim, mohammed, fairooz]
abuse	[fuck, fucking, child, shit]	[fuck, fucking, trafficker, trafficking]
cyberbullying	[bitch, fake, like, get]	[fake, bitch, quot, like]
profane	[fucktrump, fuck, dickhead, trump]	[fucktrump, dickhead, fuck, trump]
threat	[die, as, kill, going]	[die, as, kill, supertr]
harassment	[jew, fucking, whitegenocide, white]	[jew, whitegenocide, fucking, whitelivesmatter]
vulgar	[shit, hell, damn, fuck]	[shit, hell, damn, fuck]

Table 3. 4 most frequent word and 4 words with largest TF-IDF weight for each of the labels. We see that in many cases common words like ”fuck” have quite high weight in TF-IDF, that is because of their really high frequence (i.e. in label ”homophobic” there are 3346 cases of ”fucking”, 2802 of ”fuck”, third most frequent word is ”child” with just 1321 occurrences).

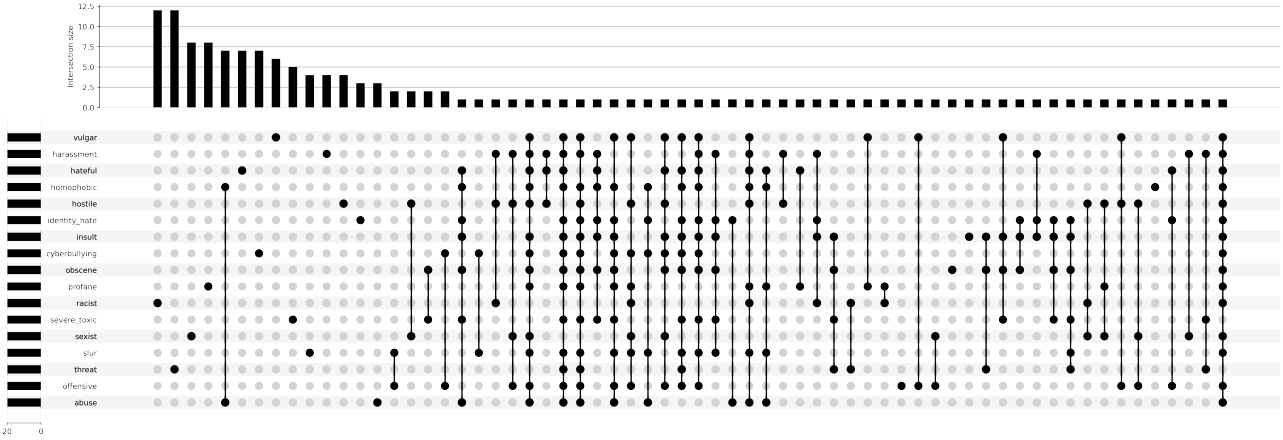

Figure 1. Upset plot for 20 most frequent words for each label, height of bar on top represents a size of intersection between labels (each label in intersection is marked with a dot below the bar). We see that labels ”racist” and ”threat” have 12 words that do not occur in any other label, while ”homophobic” and ”abuse” share 6 words between them. There are also some words that occur in many of them, we expect to get rid of such some words with use of TF-IDF

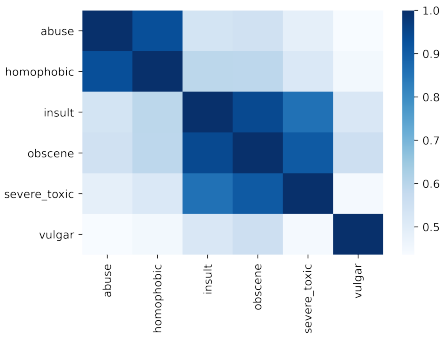

Figure 2. Cosine similarity between TF-IDF vectors of each label, we kept only the largest similarities. It seems ”abuse” and ”homophobic” are related and also ”insult”, ”obscene” and ”severe_toxic”.

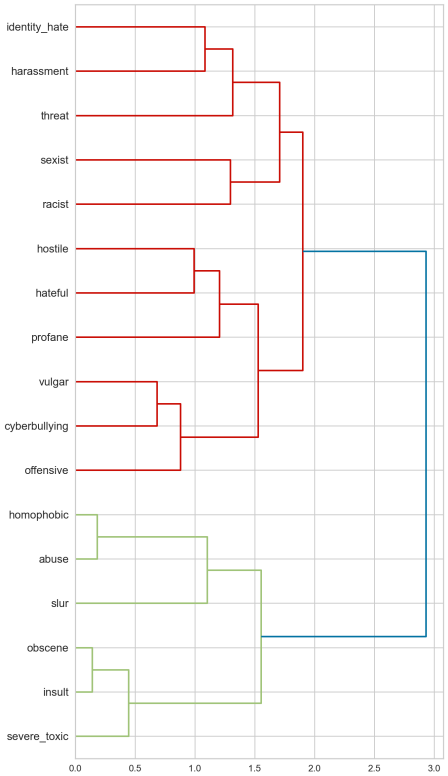

Figure 3. Hierarchial clustering done on TF-IDF representations of labels. There are few clusters visible, especially labels ”homophobic” and ”abuse” seems to be pretty similar. Same is true for ”obscene”, ”insult” and ”severe_toxic”.

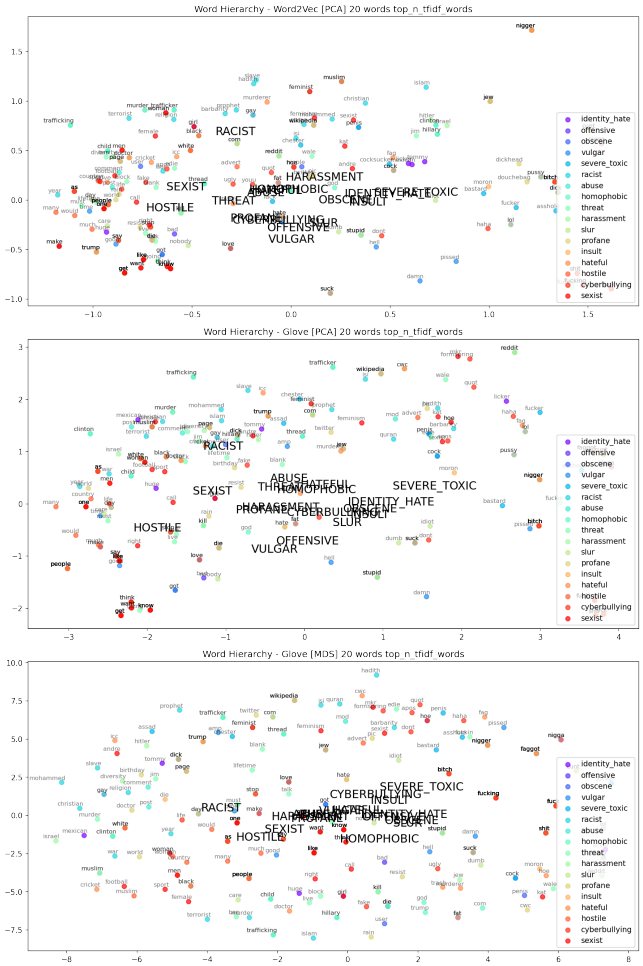

Figure 4. Pretrained embeddings and different clusterings with 20 most important words according to TF-IDF. Word2vec with PCA on the top plot, Glove with PCA in the middle and Glove with MDS on bottom.

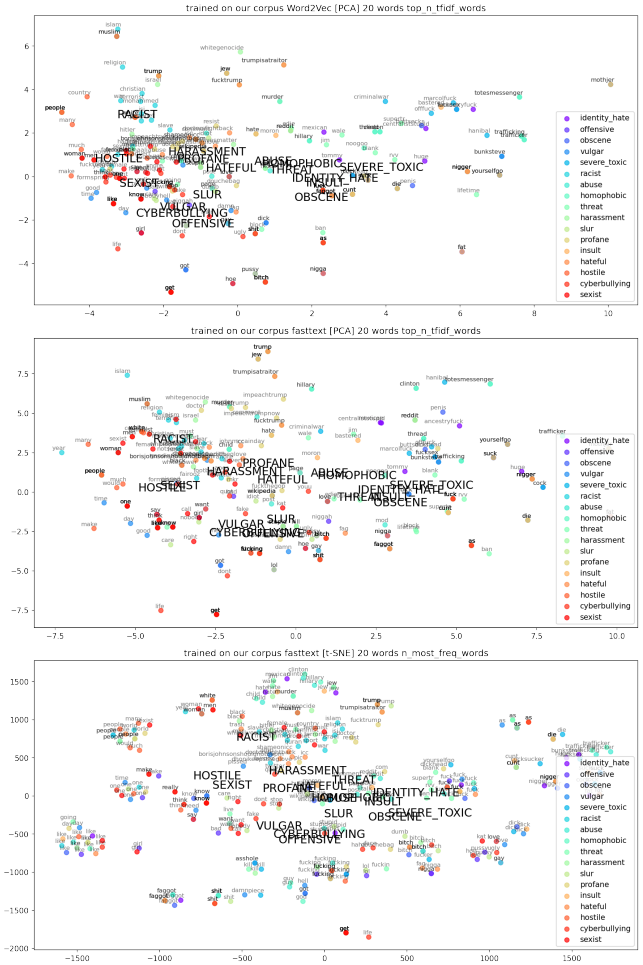

Figure 5. Embeddings trained on our corpus and different clusterings with 20 most important words according to TF-IDF or most frequent words.

we made different visualisations of clusters as described in previous section. On Figure 5 we see results we get with our own embeddings. We see that we get more clear clusters than before and which labels are similar to each other. Again we see that ”racism” label is the most isolated one similarly as ”hostile” and ”sexist”. Now we see that a ”harrasment”, ”profane” and ”hateful” could be thought of as one cluster and ”slur”, ”vulgar”, ”offensive” and ”cyberbullying” as another which makes sense as those labels are more general and not so directed such as ”racist”.

Joined labels

Based on results we got in previous sections we joined multiple labels into one: ’hostile’, ’racist’, ’sexist’ into ’HSR’; ’harassment’, ’profane’, ’hateful’ into ’HPH’; ’homophobic’, ’abuse’, ’threat’, ’severe_toxic’, ’identity_hate’ into ’HATSI’; ’vulgar’, ’slur’, ’cyberbullying’, ’offensive’ into ’VSCO’ and ’obscene’, ’insult’ into ’OI’. On Figure 6 we see results of those joined labels using some clustering techinique with two of the embeddings. As expected label ’HSR’ stands out but unfortunately we don’t get more clear cluster as we would

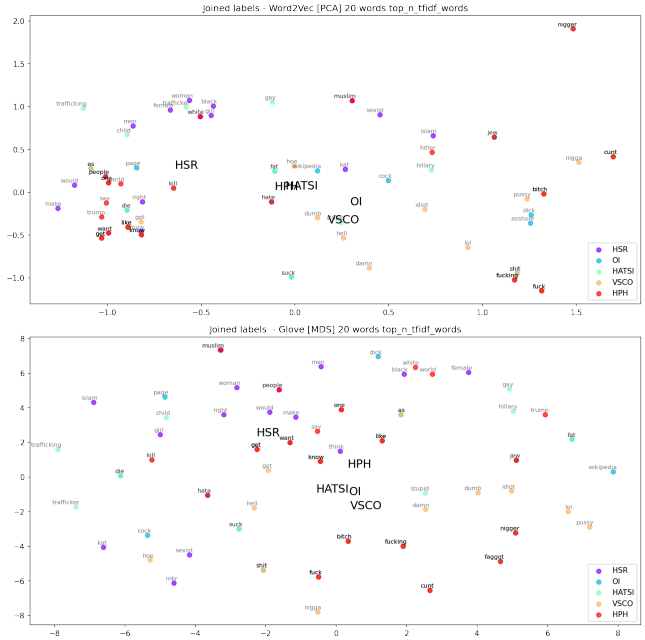

Figure 6. Joined labels with different embeddings and some clustering.

expect.

Subset of labels

We also wanted to look at only the subset of labels to see if we could get more clear picture with less noise. On Figure 7 we see an example of results we get. ”Racist” is again the most isolated and ”sexist” is closer to other labels than ”racist” which is different than before. Other labels are close to each other which would mean they are similar.

Topic modelling

Another technique we used to try to discover relationships between labels and between words is topic modelling which uses Latent Dirichlet Allocation. We did this to discover what kind of most distinct topics we would get and which words would be most representative of these topics. Number of topics to extract is a parameter and we settled for 6 topics as with more topics they were not so distinct. On Figure 8 we see relationships between discovered topics and on Figure 9 we see wordclouds and which words are the most important for each topic. If we were to try to connect some of the discovered topics with labels in our corpus, The topic 3 could easily represent the ”racism” label and topics 4 and 5 together could represent labels ”sexist” and ”homophobic”. Other topics look too general to connect them to specific labels.

Contextual embeddings

Further we also wanted to check out results we get if we use different contextual embedding techniques. For this we used 2 most common embeddings, that is Bert[17] and Elmo[18]. Since embeddings of this sort are a bit different, we needed to change our text, so for example if we had sentence ”Fuck

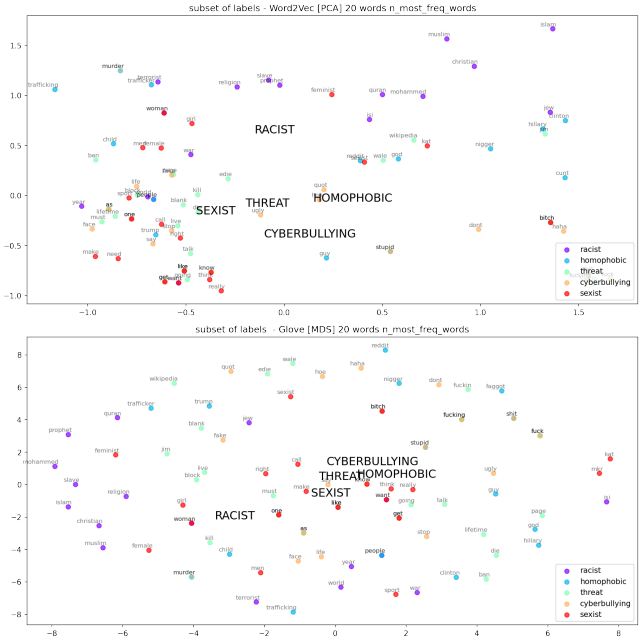

Figure 7. Subset of labels with different embeddings and some clustering.

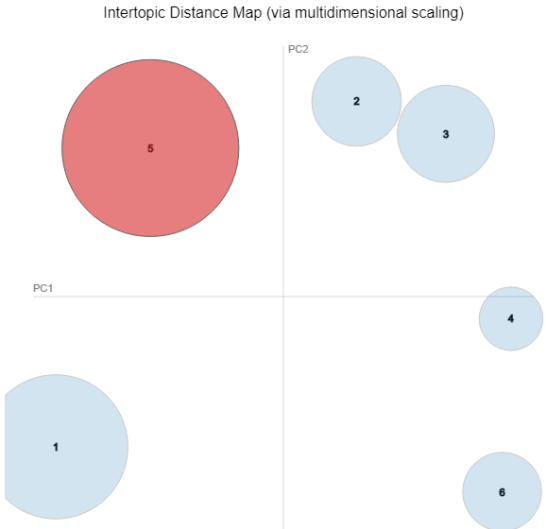

Figure 8. A representation on how close discovered topics by topic modelling (LDA) are when mapped in 2D space.

Figure 9. A visualisation of which words are most representative of some topic and how important they are (according to topic modelling with LDA).

all niggers" and its label was offensive, we changed it to "Fuck all niggers, this is offensive". In both cases we used pretrained network and did not train them on our own. Bert used was "bert-base-uncased" which was trained on huge corpus consisting of 11.038 books and English Wikipedia while for Elmo we used a model which was trained on 1 Billion Word Language Benchmark[19].

After preprocessing and tokenizing that is needed for Bert we embedded texts and then look into how word that is labeling the text is embedded. Then we randomly selected some number of texts, performed PCA and visualized results. Such picture can be seen on figure 10. Although we need to take into account that the network was pretrained and as such words already had some (pre)embbeding, we can still see that there seems to be 3 to 5 clusters of labels.

As Elmo model is capable of embedding entire sentences (if we take the output of last layer that represents a fixed mean-pooling of all contextualized word representations), here we use this as our metric for similarity. We embedded entire sentences and then performed different techniques to find out if some labels are similar to each other. For each label we took average embedded vector, performed PCA and visualized. Figure produced can be seen on figure 11. We also did something similar with Bert after we imported library that embedded sentences as whole, results are very similar to Elmo ones.

Final schema and discussion

Based on our findings described above, we were able to determine groups of labels, that share common details. Although there is no hard truth that all the techniques we used would agree to, we hand picked some clusters that occurred most

label cluster	important words
'racist'	'nigger', 'hate', 'muslim', mexican', 'white', 'black', 'islam'
'hostile', 'homophobic', 'abuse', 'sexist'	'bitch', 'pussy', 'die', 'hoe', 'sexist', 'hell'
'threat', 'severe_toxic', 'identity_hate'	'die', 'kill', 'nigger', 'fat', 'gay'
'harassment', 'profane', 'hateful'	'jew', 'white'
'vulgar', 'slur', 'cyberbullying', 'offensive'	'cunt', bitch', 'shit'
'obscene', 'insult'	'fuck', 'fucking', 'shit', 'suck'

Table 4. Final schema, we grouped labels into 6 clusters and for each cluster wrote which words are typical for it. We see that labels that are not really specific can be merged together, while racism which is specific type of hate speech stands on its own.

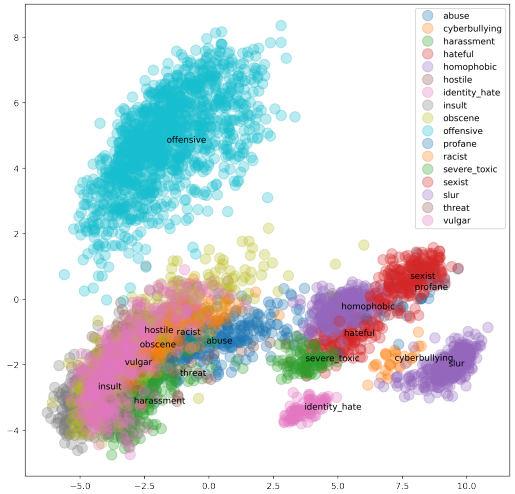

Figure 10. A visualisation of 5000 randomly selected Bert embeddings with PCA dimensionality reduction, labels "offensive" and "identity_hate" are alone, while others are in cluster with at least one another label. It is interesting to see how some labels have much more compact embeddings (e.g. "identity_hate") while others are much more spread (e.g. "offensive").

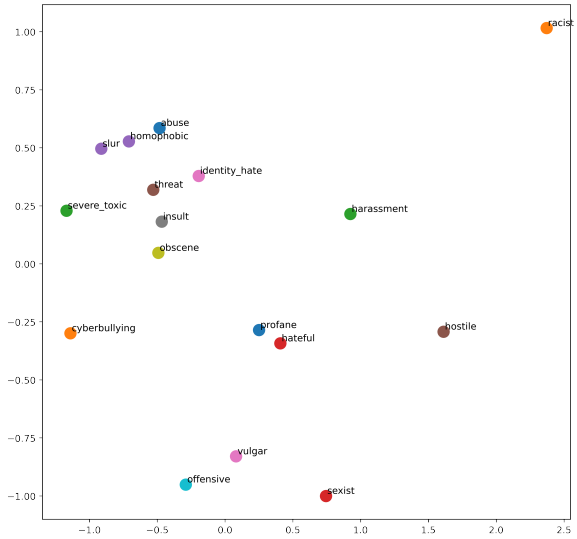

Figure 11. PCA performed on average Elmo sentence embeddings for each label (similar results are produced also with Bert sentence embeddings). Some labels ("hostile", "racist", "cyberbullying") are quite distant from the closest neighbours.

often and thus seem most reasonable. Any way we picked clusters there would be some previous work that would disagree with such clustering, but our way of choosing clusters is consistent with most of work done before. As result of our clustering we got final schema as described in table 4. In the table in the first column there are label clusters which are subsets of labels which were commonly seen or perceived as a cluster in various visualisations and could therefore be thought of as a set of synonims or as labels which could be joined together. In the second column we state a set of words which are the most representative of the labels in clusters (based on most frequent words, most important words by TF-IDF and topic modelling).

Conclusion

With our work, we were able to determine some labels that are more similar to each other. For example there is some basis to say that label "racist" speech is different from other (due to its specificity), while labels that are more often based on sex differences ("homophobic", "sexist" and "abuse") are more similar. One cluster is also composed of labels that often represent some kind of violent acts ("threat", "severe_toxic").

But all this is not really set in stone as with facts shown in Dataset section the main problem with our approach is that we are merging datasets from different sources which means that there are some identical texts that are labeled with labels, which is not optimal. That further means it is really down to the way texts are classified. For better results it would be good if we had one great dataset that is labeled by same people and unified standards.

This would also present the first step that we would do further, that is we would put more emphasis on dataset acquisition and cleaning. Also it would be interesting to manually label (on our own or even better, get specialists to do it for us) some subset of texts and then compare labels to original.

References

[1] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multi-

modal publications, 2019.

[2] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech, 2019.

[3] 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. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[4] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum, 2018.

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[6] 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.

[7] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. *IberEval@ SEPLN*, 2150:214–228, 2018.

[8] 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*, pages 1391–1399, 2017.

[9] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*, 2019.

[10] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), 2019.

[11] Susan Li. Topic modelling in python with nltk and gensim, 2018.

[12] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection, 2020.

[13] 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.

[14] Eshwar Chandrasekharan and Eric Gilbert. Hybrid approaches to detect comments violating macro norms on reddit, 2019.

[15] Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[16] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA, 2017. Association for Computing Machinery.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[19] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling, 2014.