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

# Offensive language exploratory analysis

Matic Fučka, Anže Alič and Aljaž M. Eržen

**Abstract**

In this report we present some insights into the basic construction of different sections of hate speech. We used different statistical methods to find most important words. We also used some traditional natural language processing methods and some non contextual neural approaches to finde the relationships between the classes of hate speech.

**Keywords**

Natural language processing, Statistical methods and hate speech

*Advisors: Slavko Žitnik*

## Introduction

The main objective of this assignment is to give a better understanding of the basic construction and the relationship of different subareas of offensive language.

First we constructed a dataset of hate speech with different labels with the use of different hate speech datasets which included text which were qualified under different subareas of hate speech by different reaserchers in the field. Then we preprocessed the corpus so we received the stems of each word.

After the construction of the dataset we tackled the problem of analysis by extracting the most important words for each label given to us by using some traditional natural language processing (NLP), such as TF-IDF and n-grams, methods and some statistical methods, such as Pearson $\chi^2$ test.

With the extracted words we head onto the calculation of different embeddings for these words or texts. Once we got the embeddings we tried different clustering methods such as PCA, TSNE, etc. to group these embeddings. We also discussed the results of the approaches that yielded some useful information.

## Methods

### Corpus merging

We extended our dataset selection to 7 different datasets, together including 19 different labels for hate speech types and a "not hate speech" label. Collections sources include Twitter [1, 2, 3, 4], Facebook [5], Wikipedia comments [6] and Reddit [7]. Which dataset contained which labels and how many of them can be seen in Table 2. It should be noted that some labels were present only in one dataset and some datasets had all documents labeled with a single label. Documents of one dataset were labeled with multiple labels, which we solved by mapping each of the documents into multiple replicas where each of them was labeled with one of initial labels (e.g. document labeled as `fearful_abusive_hateful` was replicated into 3 documents with labels `fearful`, `abusive`, and `hateful`).

Some datasets did not label *type of hate speech* but labeled sentiment, directness, target, group or some other property. From this we interpolated our *type of hate speech* using best effort approximations. For example, we labeled `hostile` sentiment with target `women` as `misogynistic` and target `sexual_orientation` with `homophobic`. This relabeling is not ideal due to gap between the meanings, which is why we interpolated only labels that were not found anywhere else, so there would not be clash between meanings of a single label.

Our process of merging datasets is not ideal because replicating multi-label documents introduces high correlation between labels from this dataset. Also, if labels are provided by only one dataset, we are essentially only comparing the biases of the datasets. This bias is composed of annotator's bias in interpretation of some *type of hate speech* (which is what we are after) and other biases such as collection source of the dataset or preprocessing techniques. This is why results of our analysis cannot necessarily confirm correlation between labels as correlations between *types of hate speech*, because it may be caused by other biases.

### Data preprocessing

Text in our corpus is mostly from from sources such as twitter. So we have to remove emojis, tags etc. We first replace emojis with corresponding words (e.g. replace fire emoji with word ¨fire"). Than we used special twitter tokenizer to tokenize text into tokens. We remove token which correspond to urls, mentions, unknown emojis or hashtags. On the end we stem each token with Snowball stemmer.

## Results

### Important words retrieval

We retrieved most important words in a couple of different ways. First we extracted most common 1-grams, 2-grams and 3-grams for each label. The results were mostly as expected with "retard" and "fuck" being one of the most common 1-grams for a quite high number of labels. Another interesting finding is that there is quite a lot of swear words in non-hate speech, which was quite the opposite of our expectations. We also checked which labels share the most most common words. What we noticed is that fearful, offensive, abusive and disrespectful have a lot of words, which are predominantly swears, in common. Another such cluster is toxic, obscene and insult. Other labels tend to have different words. When we looked into it we noticed these cluster have words that do make sense, so we are able to conclude that these clusters tend to cover similar topics.

Then we extracted the most statistically significant words for each label. Only the labels where it went for some sort of harassment had statistically significant words. Apperance harassment had "fatass", political harassment had "twatwaffle", and sexual harassment had "camel", "toe", "grab", "sporty", "ssi" (special sex interest) and "skullfuck". Even though we saw that each label tends to have its own most common words, there are still not a lot of statistically significant words.

We also used TF-IDF to extracted the most common words for each label. For the majority of the labels, we got good results. For example for racist we get "paki", for disrespectful "shithol", "ching" and "chong". But there are also some labels such as toxic where the most important word is "wikipedia", which does not make sense.

| Label | 3 most important words |
|---|---|
| abusive | shithol, ching, chong |
| appearance harassment | fatass, skank, rt |
| disrespectful | shithol, ching, chong |
| fearful | shithol, ching, chong |
| hateful | shithol, icc, chong |
| homophobic | dyke, game, nit |
| insult | wikipedia, jew, page |
| intelligence harassment | fucktard, rt, shithead |
| misogynistic | feminazi, mongi, chong |
| none | articl, page, wikipedia |
| obscene | wikipedia, page, edit |
| offensive | shithol, ching, chong |
| political harassment | islam, religion, twatwaffl |
| profanity | icc, doctor, mamata |
| racist | paki, beaner, sandnigg |
| sexual harassment | camel, toe, grab |
| slur | reddit, mod, com |
| threat | 0ll, supertr, wale |
| toxic | wikipedia, page, edit |
| vulgar | damn, amp, ho |

**Table 1.** 3 most important words by the TF-IDF method for each label

### Label grouping

We grouped our data using hierarchical clustering for different ways of important words extraction(n-grams and TF-IDF), then we used different types of embeddings on these words(TF-IDF, trained and pretrained FastText, Word2Vec) and clustered it different types of clusterings or dimension reductionality(PCA, LDA, MDS, T-SNE). In the section we will discuss in detail only the approaches that yielded some useful information.

Number of common word extracted with TF-IDF can be seen in figure 1. We can observe that labels abusive, disrespectful, fearful, hateful, misogynistic, and offensive are very similar. Common words are "shithol", "ching", "chong" etc. There are other similar pairs of labels such as intelligence and political harassment, appearance and sexual harassment. We come to the same conclusion if we use hierarchical clustering. The dendrogram can be seen in figure 2.

We used word embeddings from TF-IDF coefficients for the most important words for each label according to TF-IDF. 2D projection can be seen in figure 3. We can observe that appearance and intellectual harassment are in the same cluster. The same is with obscene and insult. But there are also clusters that stands for it own such as slur, threat, profanity and threat.

To project data to 2D we also used LDA (Linear Discriminant Analysis), which does not only use embeddings but also the words label. We used only the words with one label out of the most important ones. The projection can be seen in figure 4. Most important observation we can make is that the label none is separable from all of the other labels, which is to be expected. Political harassment seems differs from other labels, as it is a bit further away from the central cluster. We would probably tend to say the same if we had to judge it subjectively. Racist, vulgar and appearance harassment are very similar to each other with similar words such as "dickface".

We trained our own Word2Vec model and also use some pretrained(such as word2vec-google-news-300) ones, but the embeddings were not informative. We were not able to recognize any clusters or structure in the data. The reason is probably the fact that our corpus is not large enough, to make good estimate of the weights. Another problem is that most of data are tweets and sentences are semantically poor.

When we used the FastText model we trained it on the whole corpus but took only the labels embeddings, and reduced its dimensionality with PCA and plotted it on a 2D

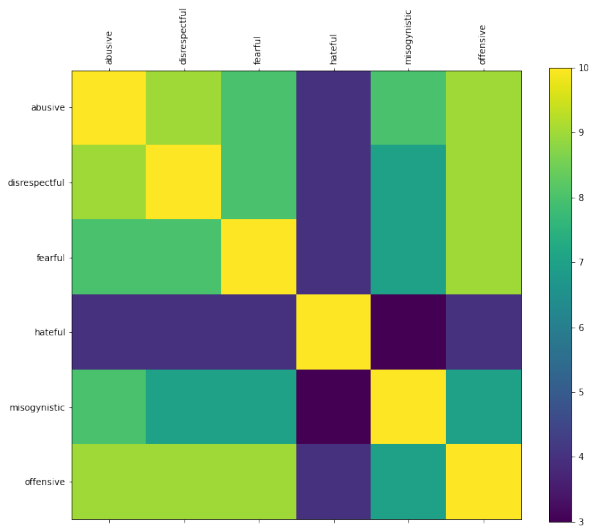


**Figure 1.** Number of common words extracted with TF-IDF.


**Figure 3.** Important words embedded with t-SNE using TF-IDF coefficients.

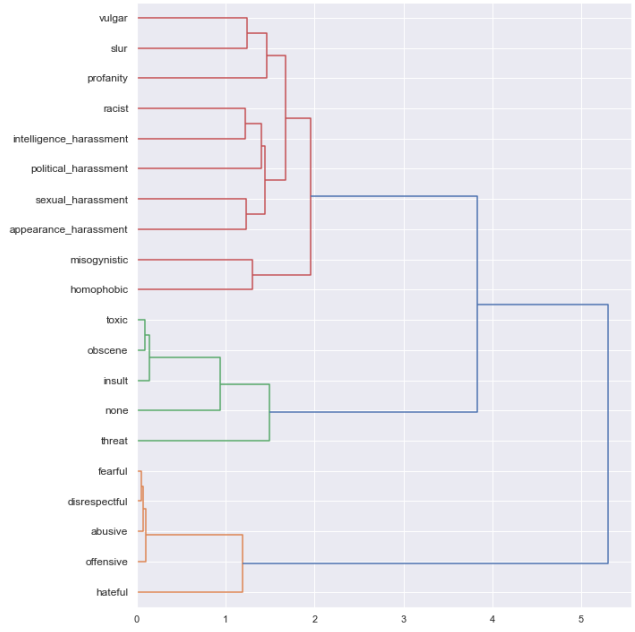


**Figure 2.** 30 important words extracted for each label with TF-IDF. Labels cluster using TF-IDF coefficients.

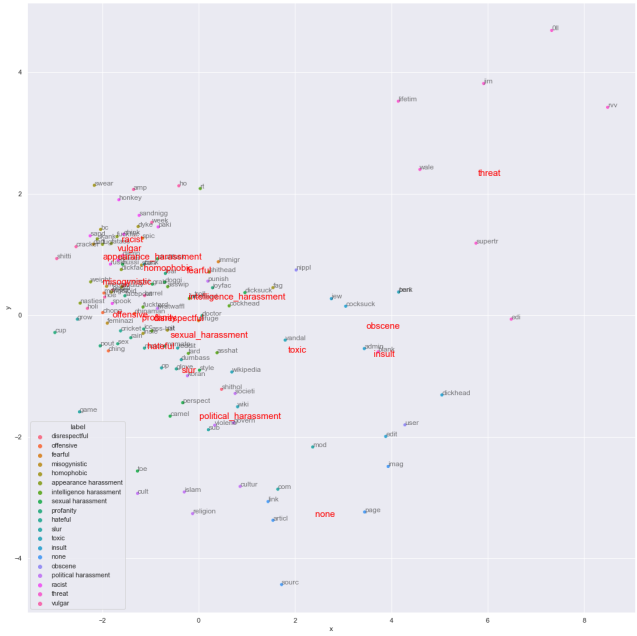


**Figure 4.** Important words embedded with FastText and projected using LDA.

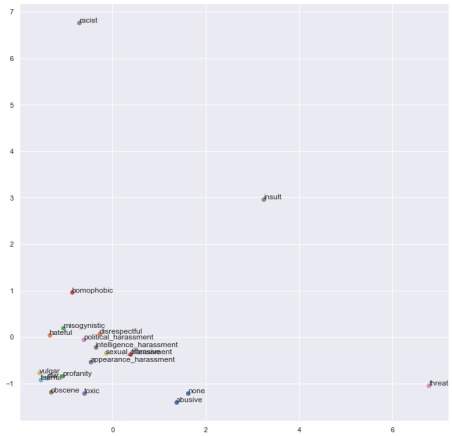


**Figure 5.** PCA decomposition of FastText embeddings for each label. FastText was trained on the merged dataset.

plane. The results are shown on Figure 5. From the figure a few clusters can be seen. As we can see we have a cluster with vulgarity, profanity, slur and fearful, which mostly makes sense as the labels tend to have a similar meaning. Another cluster is none and abusive, which does not make sense. We could explain this only by assuming that this embedding shows us some of the underlying bias of our corpus (at least for this two labels), In the centre of the coordinate system we have a slightly bigger cluster, with different types of harassment, offensive, disrespectful, hateful and misogynistic. The common line with this labels is that they tend to focus on the ”outside” or the ”inside” properties of the victim. Homophobic, racist, insult and threat are laying outside of these clusters.

FastText was also used to find most similar words to labels. Embedding can be seen in figure 6. For some labels we get exquisite results. For instance, with the label racist we get that some of the most similar words are ”neonazi”,”antisemit”, ”facist” etc. But for many other labels most similar words do not make a lot of sense. Most of the labels are separable from the others and forms compact clusters such as slur, cyberbullying, vulgar, hateful, obscene etc. But there are labels which are very similar. For instance threat, abusive, offensive. For the compact ones we are able to say that most probably these labels cover their own area of swear words.

## Future work

In the future we will use embeddings of the most important words for each labels, returned to us by the last layer of different pretrained BERT[8, 9] and ELMO[10]. With these embeddings we will try different methods of clustering and from all of the information we gathered through all of our experiments and tried to make a schema of the labels [11, 12] we have in our corpus and the schema for all of the prescribed labels using similar approaches such as.

## References

[1] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. https://github.com/HKUST-KnowComp/MLMA_hate_speech, 2019.

[2] Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit P. Sheth. Publishing a quality context-aware annotated corpus and lexicon for harassment research. http://arxiv.org/abs/1802.09416, 2018.

[3] 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. http://aclweb.org/anthology/C18-1248, 2018.

[4] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Hate speech and offensive content identification in indo-european languages. https://doi.org/10.1145/3368567.3368584, 2019.

[5] 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. https://github.com/marcoguerini/CONAN, 2019.

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

[7] Eshwar Chandrasekharan and Eric Gilbert. Hybrid Approaches to Detect Comments Violating Macro Norms on Reddit. https://zenodo.figshare.com/articles/dataset/Hybrid_Approaches_to_Detect_Comments_Violating_Macro_Norms_on_Reddit/7598930, 1 2019.

[8] Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. Toxccin: Toxic content classification with interpretability, 2021.

[9] Suman Dowlagar and Radhika Mamidi. Hasocone@firehasoc2020: Using bert and multilingual bert models for hate speech detection, 2021.

[10] Michal Bojkovský and Matúš Pikuliak. STUFIIT at SemEval-2019 task 5: Multilingual hate speech detection on Twitter with MUSE and ELMo embeddings. 2019.

[11] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, 2018.

[12] David Andrzejewski and Xiaojin Zhu. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for NLP*, pages 43–48, 2009.

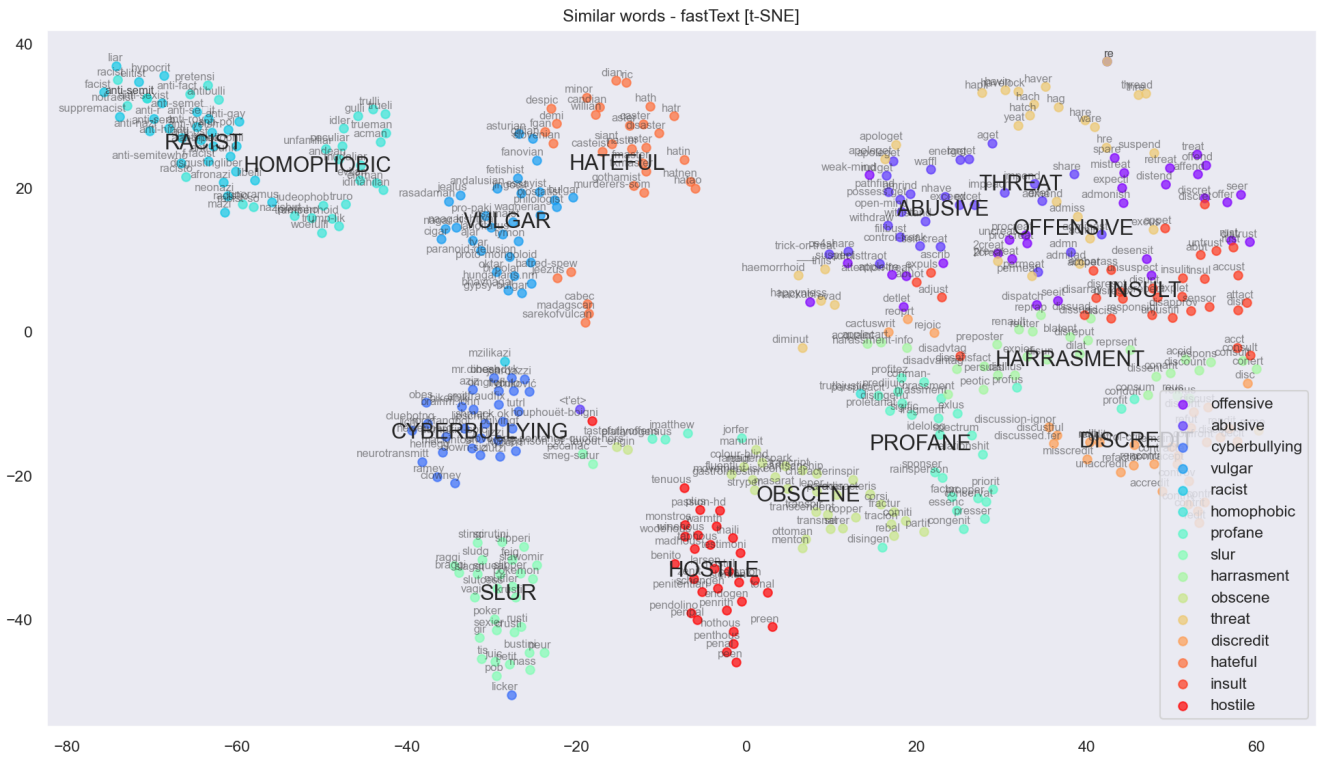


**Figure 6.** Most similar words retrieved by fastText.

| Sentiment | 16 fcbk | 20 twtr | 25 twtr | 32 twtr | wikipedia | vulgar twtr | reddit | total |
|---|---|---|---|---|---|---|---|---|
| abusive | 0 | 671 | 0 | 0 | 0 | 0 | 0 | 671 |
| appearance_harassment | 0 | 0 | 0 | 677 | 0 | 0 | 0 | 677 |
| disrespectful | 0 | 782 | 0 | 0 | 0 | 0 | 0 | 782 |
| fearful | 0 | 562 | 0 | 0 | 0 | 0 | 0 | 562 |
| hateful | 0 | 1278 | 1267 | 0 | 0 | 0 | 0 | 2545 |
| homophobic | 0 | 845 | 0 | 0 | 0 | 0 | 0 | 845 |
| insult | 0 | 0 | 0 | 0 | 7877 | 0 | 0 | 7877 |
| intelligence_harassment | 0 | 0 | 0 | 810 | 0 | 0 | 0 | 1360 |
| misogynistic | 0 | 1360 | 0 | 0 | 0 | 0 | 0 | 33157 |
| none | 0 | 0 | 4456 | 21059 | 7642 | 0 | 0 | 810 |
| obscene | 0 | 0 | 0 | 0 | 8449 | 0 | 0 | 8449 |
| offensive | 0 | 4020 | 522 | 0 | 0 | 0 | 0 | 4542 |
| political_harassment | 3864 | 0 | 0 | 698 | 0 | 0 | 0 | 4562 |
| profanity | 0 | 0 | 760 | 0 | 0 | 0 | 0 | 760 |
| racist | 0 | 0 | 0 | 702 | 0 | 0 | 0 | 702 |
| sexual_harassment | 0 | 0 | 0 | 229 | 0 | 0 | 0 | 229 |
| slur | 0 | 0 | 0 | 0 | 0 | 0 | 5059 | 5059 |
| threat | 0 | 0 | 0 | 0 | 478 | 0 | 0 | 478 |
| toxic | 0 | 0 | 0 | 0 | 16889 | 0 | 0 | 16889 |
| vulgar | 0 | 0 | 0 | 0 | 0 | 6718 | 0 | 6718 |
| total | 3864 | 9518 | 7005 | 24175 | 41335 | 6718 | 5059 | 97674 |

**Table 2.** Number of documents for different sentiments and sources. Prefix numbers of sources are indexes from https://hatespeechdata.com/