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.

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 ext 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 neiher of those. They use this dataset to build features such as bigram, unigram, triram TF-IDF, and they also include features such as length of tweet, number of hashtags, mentions, retweets,... They then test a variety of models t: 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 in 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 entry is targeted or not. For 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 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.

Initial Ideas

- Combine and preprocess datasets so they have the same (or very similar) structure so we will be able to compare them.
- Compute and visualise word and phrase statistics, and how are they correlated with hate speech and different types of hate speech.
- Make different features for posts (tweets): bag of words, TF-IDF on unigrams, bigrams, trigrams. Try to improve with dense embeddings.
- Make clustering on words or posts to see if we can get clear clusters which we would then compare with actual labels of hate speech types. Also visualise which words are similar by measuring the distance between them.

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 https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data 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).
- Vudgen 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

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.

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 (originally provided at https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection) 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).

- Golbeck et al. [16] prepared a large human-labeled corpus for online harassment research where they labeled 35000 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 almost every 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 then clusters already seen on few more clusters are visible (”vulgar”, ”cyberbullying” and ”offensive”).

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, MDA and TSNE, one such visualization can be seen on figure 4. Another example using FastText and t-SNE can be seen on figure 5.

When using PCA for dimensionality reduction (it really does not matter, which words and technique we choose to

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.

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).

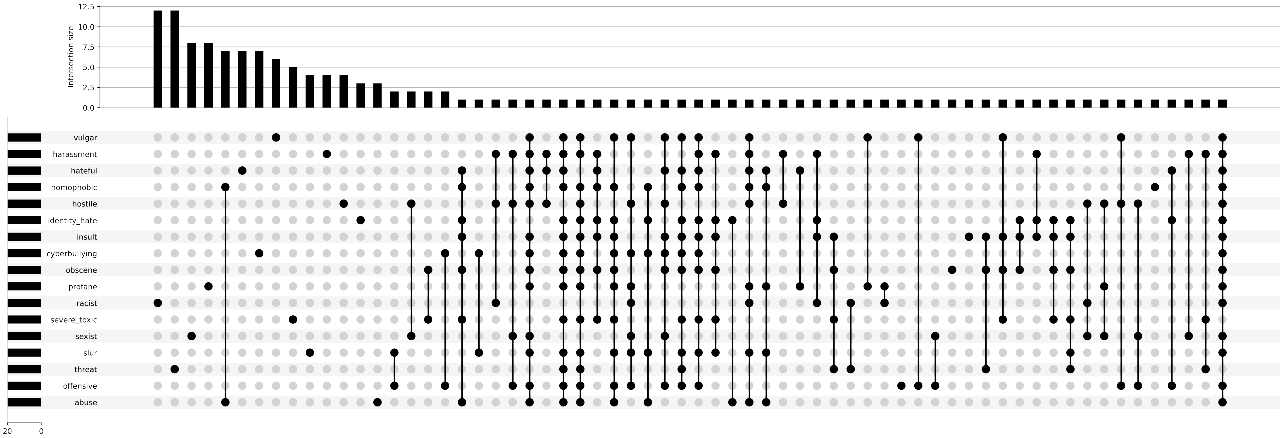

Figure 1. Upset plot for 20 most frequent words for each label, height of bar on top represents a size of intersection between labels (each label in intersection is marked with a dot below the bar). We see that labels "racist" and "threat" have 12 words that do not occur in any other label, while "homophobic" 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.

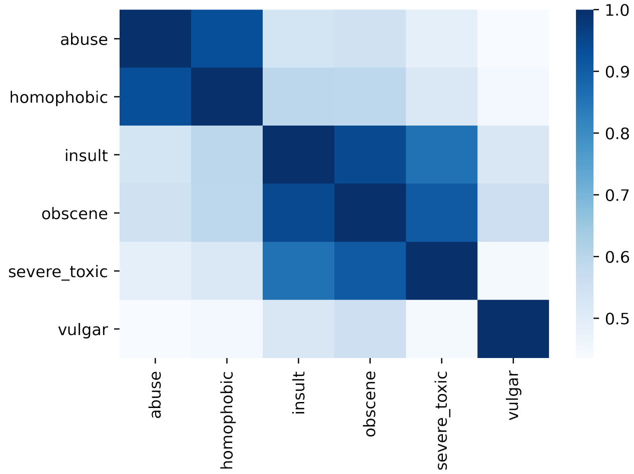

Figure 2. Cosine similarity between TF-IDF vectors of each label, we kept only the largest similarities. It seems ”abuse” and ”homophobic” are related and also ”insult”, ”obscene” and ”severe_toxic”.

embed) most of the labels are clustered close together, but there are some exceptions. Especially label ”racist” is quite distant from the closest neighbour, and also labels ”sexist”, ”hostile” and ”profane” looks like they do not belong to same cluster as most of the labels. On the other hand labels ”threat” and ”hateful” seems like they are almost the same, as is the case with ”severe_toxic” and ”obscene”

Another thing we tried is that we used word2vec to embed each word of our documents (tweets) and then weighted each vector corresponding to TF-IDF weight of selected word. For document embedding we calculated just the average vector of each word of selected document. We then further calculated also average vector for each label and then used PCA to visualize labels in 2D space. Such representation can be seen on figure 6.

Here it once again seems that labels ”racist”, ”hostile” and ”sexist” belong on their own, while some labels are close together (see ”insult” and ”obscene”). Some labels that are distant from others in this case are also ”slur”, ”offensive” and ”severe_toxic”.

So far we have seem to establish some results:

- label ”racist” is not similar to any other label,
- labels ”insult” and ”obscene” are somehow connected (in some interpretations label ”severe_toxic” can be added to them),
- same is true for labels ”homophobic” and ”abuse”,

of course these are all just first propositions and we will try to establish more firm results in the future.

Future work

- train used models (word2vec, fastText and GloVe) on our data,

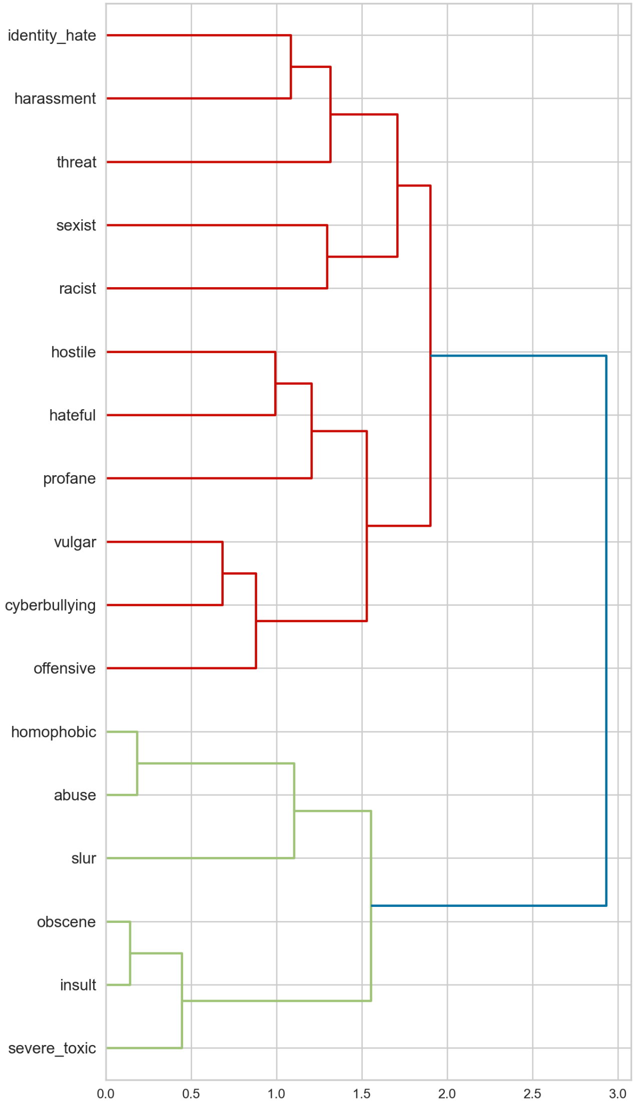

Figure 3. Hierarchial clustering done on TF-IDF representations of labels. Few clusters are formed.

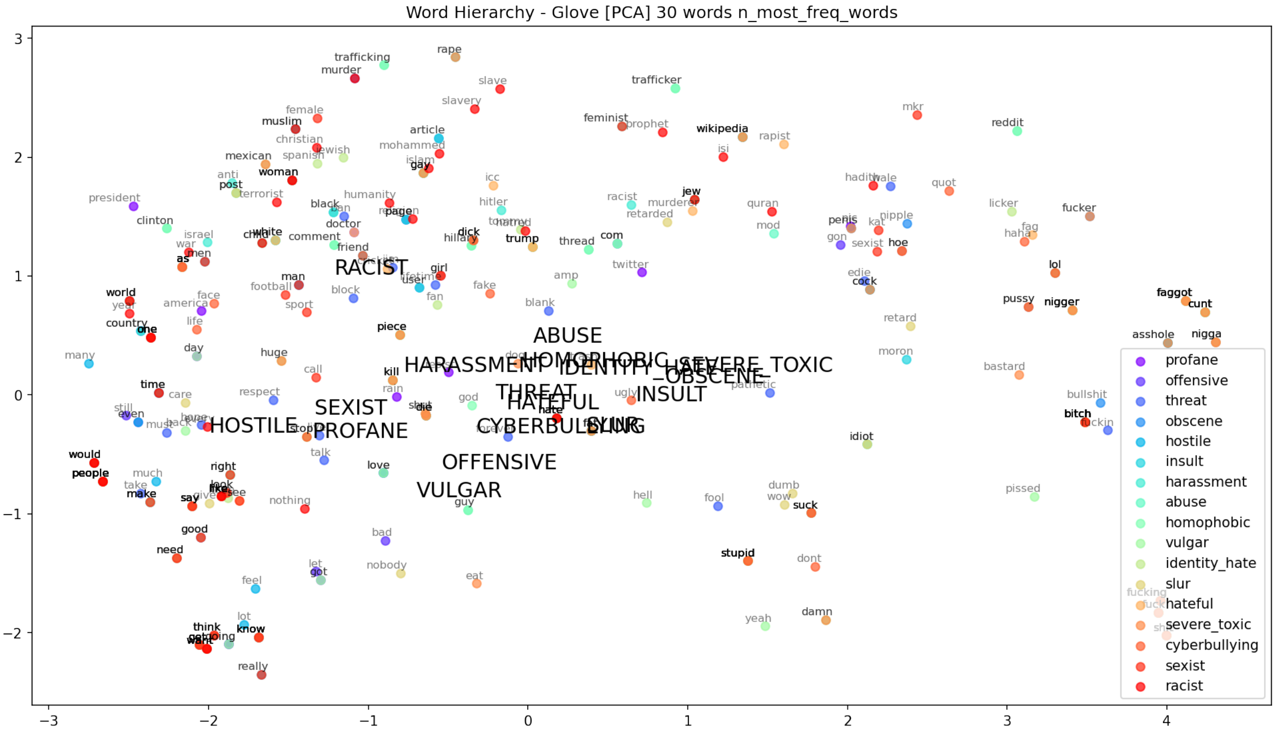

Figure 4. PCA done on GloVe representations of labels (we used 30 most frequent words for each label).

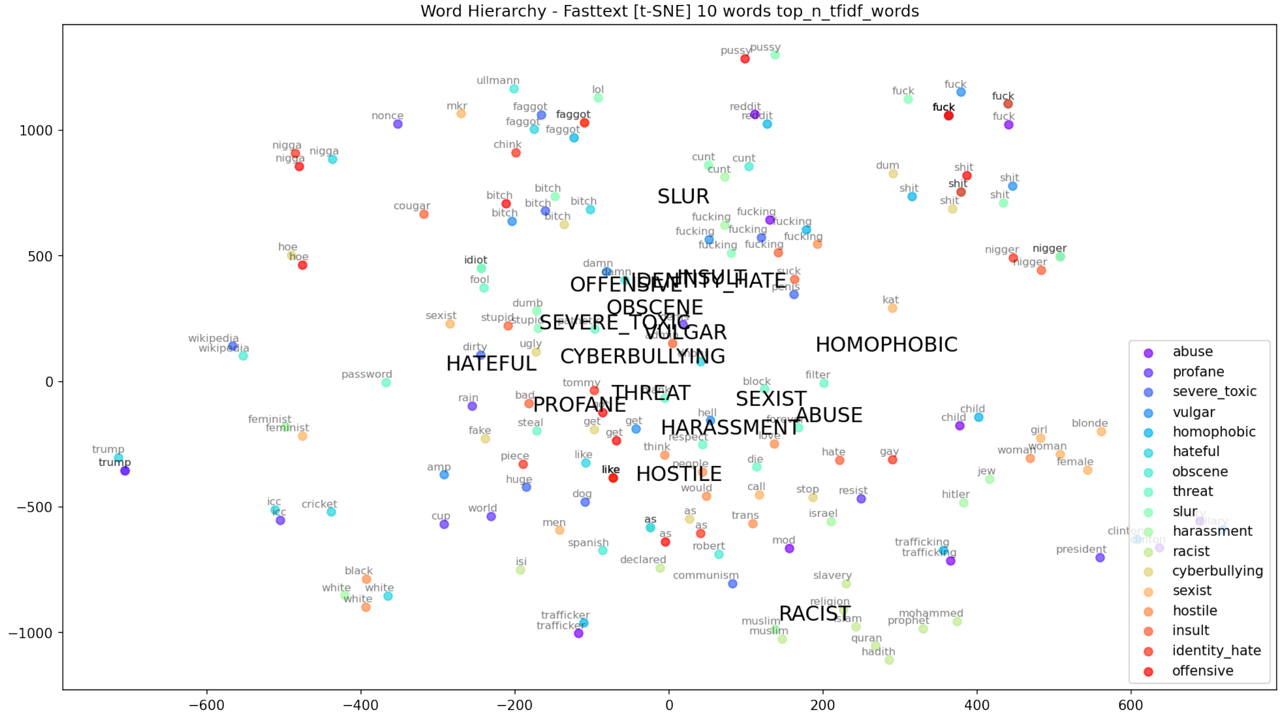

Figure 5. t-SNE done on FastText representations of labels (we used 10 most important words by their TF-IDF weights for each label). We observe ”racism” is its own cluster.

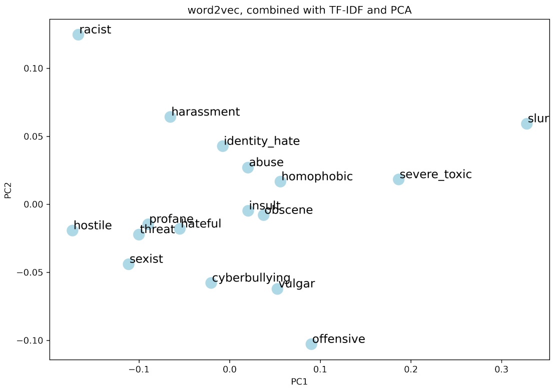

Figure 6. Combination between word2vec and TF-IDF. ”racist” label is once again on its own.

- use contextual word embeddings (Bert and Elmo),
- try to use some topic modelling techniques (i.e. Latent Dirichlet allocation) to see if we can discover some new connections between labels
- define final schema for offensive language

References

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