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

Offensive language exploratory analysis

Benjamin Džubur, Žan Magerl and Jurij Nastran

Abstract

TBA

Keywords
Offensive language, clustering, visualizations, embeddings

Introduction

The use of offensive language and various offensive terms are often labeled with one or multiple of different categories. These categories are e.g. sexism, racism, hate speech, misogony, derailing, sexual harrasment, religion based, islamophobia, terrorism, etc. Notice how some of these categories are subcategories of others, some may be very closesly related or perhaps addressing the same issue. When data is labelled with some of these categories, it is usually done subjectively to a degree, and not by professional linguists. Therefore, the idea of this project is to use NLP and ML methods to analyze, explain and compare different categorizations of data, see what they consist of and how they relate to each other.

Related work

Many similar approaches have been performed on this and similar topics. Some notable are:

- Jaki & De Smedt [1] used skip-grams and k-means clustering on the top 250 most biased words of German hate speech tweets to identify clusters of words that are related.
- Rodriguez et al. [2] used TF-IDF representations of Facebook posts and related comments, which they then used with k-means clustering and extracted the most related words of the clusters and analyzed the topics.
- Reimers et al. [3] used agglomerative hierarchical clustering, k-means and DBSCAN clustering algorithms to cluster most similiar arguments. They studied how they can improve sentence-level argument similarity and clustering by using contextualized word embeddings.
- Sia et al. [4] compared two centroid based clustering algorithms on different word embeddings (Word2vec, ELMo, BERT, Fasttext etc.) in task of extracting top 10 words for different topics from news datasets.
- Malmasi et al. [5] used skip word bigrams and n-grams for word representation and then used Brown clustering in order to discriminate between hate speech and offensive speech.
- Ayo et al. [6] implemented a naïve Bayes model to improve feature representation and proceeded to use a modified Jaccard similarity measure for real-time probabilistic clustering of tweets into hate speech topic clusters.

Methodology

In order to use ML approaches such as clustering, we must represent words and documents numerically. We use different types of embeddings for this. A simple sparse embedding method is generating a TF-IDF matrix, using which we get simple representations for both words and documents. Another option is using neural, pre-trained word embeddings such as Word2vec to represent words, and document embeddings may be calculated based on averages of the containing words. For contextual embeddings such as BERT and ELMo, we can pass sentences of documents as an input.

In order to group the most related words, we choose to use K-Means clustering, Agglomerative hierarchical clustering, DBSCAN and Affinity propagation because of their notably differing approaches. We analyse the results of the clustering methods on different kinds of embeddings and provide insights into how the choice of embedding effects the resulting clusters. We use measures such as homogeneity and V-measure score to evaluate the sensibility of categories.

For dimensionality reduction, we will resort to the unsupervised t-SNE and PCA, as well as the supervised LDA, as they have proven to be the methods of choice.

In the Experiments section, we go into further detail into how exactly the above methods were used.

Data gathering & preprocessing

In the data gathering phase we've collected texts from 9 different datasets. Reference to each dataset and labels that we've used in our merged dataset can be found in the Table 1.

Most of the datasets have contained tweets, but there were some exceptions. In the following list we've pointed out some information about used datasets:

- Dataset from Davidson et al. [7] contains manually labeled Twitter tweets using CrowdFlower workers.
- Dataset from Reynolds et al. [8] contains posts from social Q&A platform Formspring.me that is notorius for many cyberbullying and harrassment incidents. Each extracted post was labeled by three independent workers from Amazon Mechanical Turk.
- Dataset from Founta et al. [9] contains manually labeled Twitter tweets using CrowdFlower workers.
- Dataset from Mandl et al. [10] contains Twitter tweets and Facebook posts that were labeled with selected group of workers through their own annotation system.
- Dataset from Cachola et al. [11] contains Twitter tweets, that were sampled from those posted by a set of users with known socio-demographic traits obtained through asking the users to self-report them in an online survey.
- Dataset from Kaggle [12] contains Wikipedia comments, that were labeled by human reviewers for toxic behavior.
- Dataset from Fersini et al. [13] was dataset used in Automatic Misogyny Identification task in IberEval evaluation campaign. It contains Twitter tweets labeled by CrowdFlower workers.
- Dataset from Gomez et al. [14] is multimodal dataset that contains tweets and was annotated using the crowdsourcing platform Amazon Mechanical Turk.
- Dataset from Jha et al.[15] contains manually labeled Twitter tweets.

In the Figure 1 we can see distribution of data categories in our corpus. We should note that we've used the alternative names of some of the categories in the future figures and tables in order to make the analysis more coherent. Category *hostile sexism* was renamed into *hostile*, *benevolent sexism* into *benevolent* and *sexual_harrasment* into *harrasment*.

Preprocessing

In every text we've removed punctuations and splitted sentences in the sequences of words. After that the next step was to remove common English stopwords. With the help of regular expression we've also removed tags of other users (commonly denoted with @), tokens not containing any letters (emojis, numbers), links to other pages. Some texts have already anonymised data and have therefore replaced user tags with labels such as `<USER>`. In this step we have also removed these kind of labels. At the end we've have performed lemmatisation on the remaining words.

Sources	Categories
Davidson et al. [7]	offensive
Reynolds et al. [8]	cyberbullying
Founta et al. [9]	hateful, abusive, spam
Mandl et al. [10]	profane
Cachola et al. [11]	vulgar
Kaggle [12]	obscene, threat, insult
Fersini et al. [13]	derailing, discredit, dominance, sexual harrasment, stereotype
Gomez et al. [14]	homophobic, racist
Jha et al. [15]	benevolent sexism, hostile sexism

Table 1. Dataset sources and list of keywords that we obtained from each dataset.

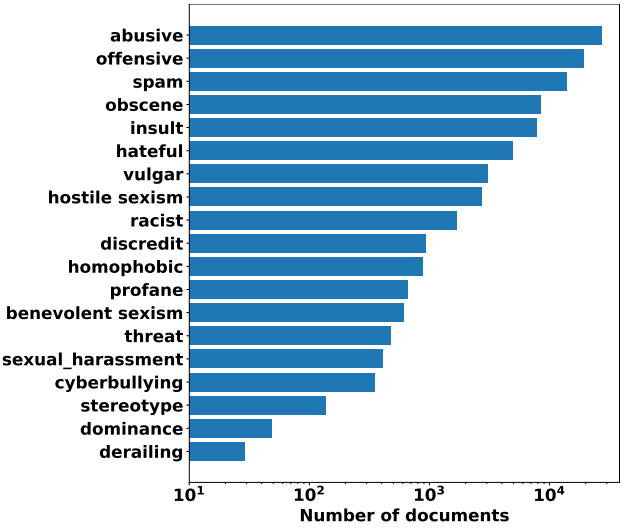

Figure 1. Data distribution

Data exploration

In order to get more detailed insight into data we've also delved into data exploration.

First we've look into the length of texts for different categories. For every category we've calculated the average length of its texts and we've immediately observed three outliers: texts labelled with *insult, obscene* or *threat* were significantly longer (300 words) than other texts (around 100 words). This is due to the origin of the texts, since these texts originated from Wikipedia comments opposed to the most other datasets which were composed mostly from tweets, which are of the restricted length (280 characters, i.e. 80-100 words).

Some of the texts from our dataset contain very interesting sequences, some of which are listed here:

- Wikipedia comment (*obscene*): *poop. poop. poop. poop. poop. poop. poop. poop. poop. poop. poop.*

poop. poop. poop. ...

- Tweet (*spam*): *I'm trying to win a FREE Nintendo Switch from @EsportsArena! You can win one too!*
- Formsping.me post (*cyberbullying*): *no body likes yuuuuuuu!*
- Tweet (*profane*): *@robreiner He is so fucking stupid. #FuckTrump*
- Tweet (*benevolent*): *Appreciate your lady, your mum, your sister today and every day #WomensDay*

We can see that some tweets are very time/topic specific. For example, some of the most common words from tweets from the dataset described in Mandl et al. [10] were associated with the former US president Donald Trump. Tweets are from second half of 2019 and could be linked to his first impeachment. Another example are the *benevolent* tweets, whose content is commonly associated with the women's day.

Experiments & Results

TF-IDF based representation

Our first pipeline resorted to TF-IDF representation of documents. We sampled a maximum of 500 documents from each category for a balanced representation and to simplify the clustering output. After preprocessing, we built the TF-IDF matrix on unigrams. Based on the mean TF-IDF weights of documents of each category, we extracted the keywords, which are shown in 2.

Table 2. Top keywords of each category, based on TF-IDF scores.

category	keywords
abusive	fucking,fucked,bad,ass,like,cant,idiot,hate
benevolent	man,woman,love,womansday,like,women,girl
cyberbullying	bra,get,know,bitch,dont,fake,like,fuck
derailing	women,rape,woman,men,lol,bitch, whore
discredit	bitch,stupid,whore,hoe,girl,slut
dominance	bitch,women,like,whore,men,yesallmen
hateful	niggas,hate,nigga,fucking,idiot,like,mad
homophobic	faggot,dyke,see,look,like,called,gay,straight
hostile	sexist,mkr,women,kat,notsexist,girls
insult	fuck,fucking,go,wikipedia,shit,page
obscene	fuck,fucking,dont,go,suck,get,shit
offensive	bitch,bitches,hoes,pussy,hoe,like,fuck
profane	fucktrump,dickhead,douchebag,trump,fuck
racist	nigger,nigga,white,trash,like,black,card
harrasment	bitch,dick,rape,girl,cock,ass,pussy,fuck,suck
spam	new,free,amp,video,via,win,check,enter
stereotype	women,dick,woman,bitch,girl,like,whore
threat	fuck,fucking,go,wikipedia,like,stupid
vulgar	shit,hell,fuck,fucking,bitch,ass,damn,pissed

At first glance, documents from many different categories tend to include the same keywords. Some categories immediately stand out however: namely *profane*, *spam* and *benevolent*. For *profane*, this is expected as this is the only category from one dataset, for which the tweets were gathered from the political domain. Tweets tagged as *benevolent (sexism)* use well-meant and kind words, so it is logical for it to be distinct from others. Similarly, for *spam*, the content usually includes promotions and is not offensive.

We performed K-Means clustering with K equal to the number of all categories (19) on the TF-IDF representations of documents in order to find whether or not the categories made much sense. We evaluated how well a clustering result fits each individual category via the V-measure, which is a harmonic mean of homogeneity score (here we want each cluster to contain only members of a single class) and completeness score (we want all members of a given class to be assigned to the same cluster). Based on this we found that the categories that made the most sense to be distinct were (in this order): ***homophobic, racist, benevolent sexism, hostile sexism, abusive, profane, spam*** . For these categories, we could easily find predicted clusters which correspond to them, as the top keywords were almost identical. The other categories had very low V-measure score, meaning the documents of those categories had very diverse cluster predictions, and that these categories could be somehow related.

Finally, we can look at the plot 2, which nicely confirms our above statements. We first used PCA to reduce dimensionality of our TF-IDF matrix to 50 dimensions, which we then further used as an input to t-SNE, which reduced the data to 2 dimensions. We annotated some notable clusters of high enough homogeneity to improve readability. We can clearly see that the clusters we pointed out correspond to the categories which we previously found sensible to be distinct. Furthermore, we found that the centroids of those other categories were quite similar in this 2D space.

Additionally, we implemented Linear Discriminant Analysis (LDA) to find a linear transformation that best separates the categories. The results, shown on 3, are less impressive than the approach with T-SNE, successfully separating only instances from the categories 'benevolent sexism', 'homophobic' and 'profane', while the other categories are are mostly bunched together around the $(0, 0)$ point.

TF-IDF weighted average Word2vec representation

We used the pre-trained Word2vec model on Google news as a tool to extract document embeddings. We then used those to perform clustering on categories and determine their similarities. We used two slightly different approaches here.

Firstly, using the TF-IDF representations of documents as weights, we computed the average Word2vec representation of each document. For each category, we then computed the mean vector by averaging the Word2vec vectors of the corresponding documents. We computed cosine similarities between these category vectors, and performed affinity propa-

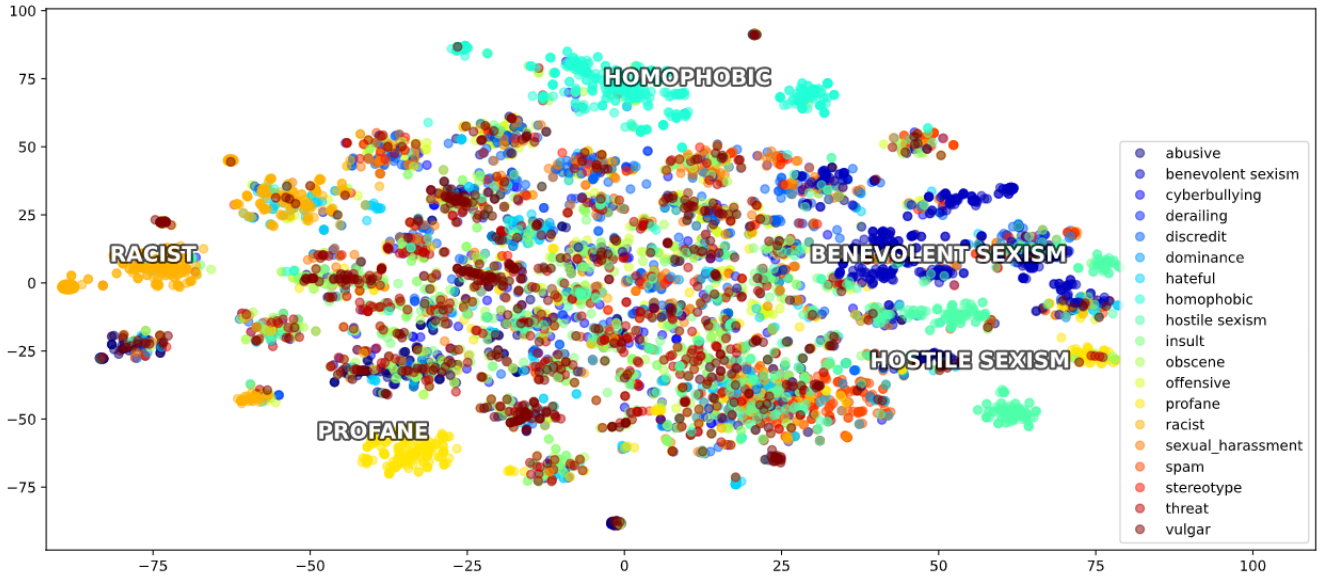

Figure 2. Document embedding visualization, true categories are encoded via color. T-SNE was used for dimensionality reduction. Some notable homogeneous clusters are annotated.

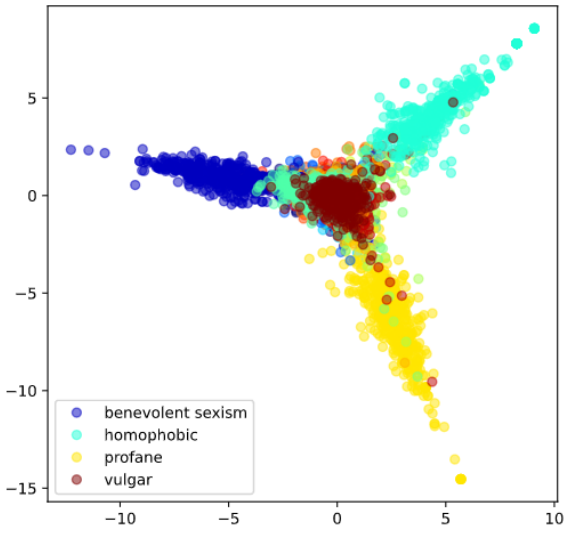

Figure 3. LDA result in 2 dimensions with only the four most noticeable categories annotated.

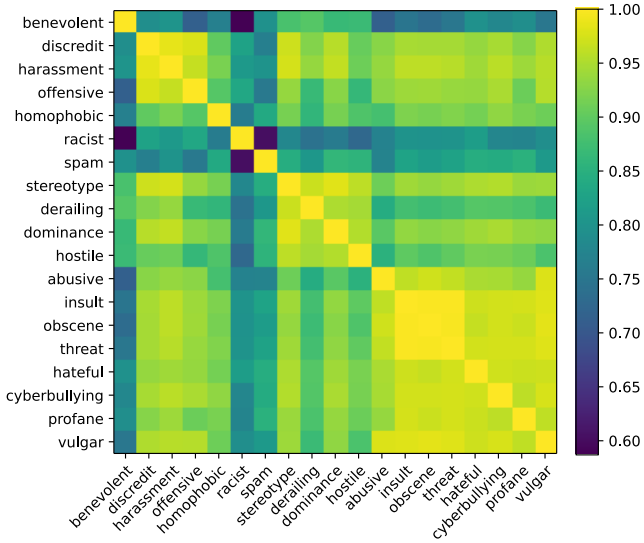

Figure 4. Similarities between categories based on TF-IDF weighted Word2vec representation of documents.

gation clustering before plotting the similarity matrix in figure 4. We immediately notice that categories *insult, obscene, threat* are used to label very similar documents. We confirm the similarities between these categories by looking at their representative keywords and after acknowledging that these three categories come from the same dataset. The similar holds for categories *stereotype,derailing,dominance*, which also come from one dataset with 5 categories.

For the second approach, we merged documents of the same category together and then used only the TF-IDF representations of the 19 resulting documents, from which we computed 19 average Word2vec representations. We then calculated the cosine distance matrix with those vectors and performed agglomerative hierarchical clustering with complete-linkage. The resulting dendrogram is shown in figure 5. It is worth noting that the resulting hierarchy differs considerably depending on the used linkage. *Harassment* and *discredit* categories are shown to be most similar in this case with *threat, insult* and *obscene* being a notably similar trio, which seems reasonable, both intuitively and sue to the fact that the three categories were also found to be similar in figure 4. On the other hand, categories *hateful* and *cyberbullying* seem to be similar in figure 4, but are put in different places in the hierarchy in the case of complete-linkage and are therefore an

example of results that differ depending on the method used.

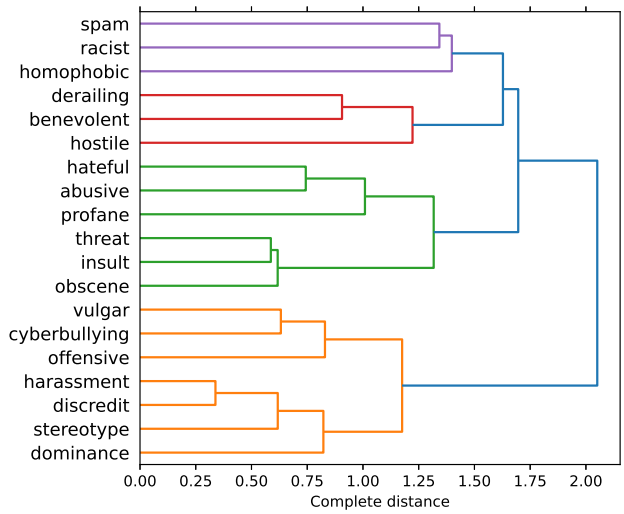

Figure 5. A hierarchical representation of inter-category distances using complete linkage.

Exploration of terms via Word2vec

We were given a set of terms which we should explore the meaning of, without necessarily having data, categorized by these terms. In this pipeline we again used Word2vec embeddings for these terms. We've focused on two approaches:

- we've compared Word2vec vectors of the terms with each other
- we've compared Word2vec vectors of the terms with average Word2vec vectors of our data categories from our dataset

In both approaches we've used the Word2vec model that was pre-trained on Google news.

In our first approach we've extracted Word2vec vector for every term. Then we've computed similarity between every pair of vectors and constructed the similarity matrix. After that we've applied the affinity propagation clustering and obtained 5 different clusters. We've found that that the terms *vulgar, profane, obscene* are in Word2vec space very similar. Similar holds for the terms *racist, homophobic, hostile*. The term that differed the most from other terms was the term *threat*, which is probably due to the more direct/explicit vocabulary, such as *killing, threat, danger, harm*, commonly used in such texts, which is not present in other categories.

One of the advantages of this approach is that we can extract some information about the the terms that we're not able to obtain in our dataset, which was not possible with the previous methods. In our case is this the term *slur*, which is most similar to the category racist. Therefore we can believe that there is linkage between racist and slur and we will try to confirm this in our second approach.

In our second approach we've have first obtained Word2vec vectors for our terms, similarly to the first approach. To obtain average Word2vec vectors of our data categories we've used similar procedure as in the previous pipeline, where we've first merged all documents of the same category into one document, therefore obtaining 19 documents. Then we've built TF-IDF matrix and used its weights in order to compute average Word2vec vector for every document/category, from words that were present in the Word2vec model. The only remaining step to do was to compute the cosine similarity between every term Word2vec vector and every average Word2vec vector from categories. The result can be presented as 19×15 similarity matrix than is visualised in the Figure 6.

We immediately notice that the terms *offensive, cyberbullying, threat, discredit, hostile* are visually separated from the rest. That means that the Word2vec vectors obtained from the our model do not well match our data.

Similarly to the first approach we can with this approach also extract some information about the the terms that we're not able to obtain in our dataset. And according to figure 6, *slur* is most similar to the *racist* texts in our dataset. This can be seen as confirmation of our observation in the previous approach.

Another observation that we can make is that the data category *spam* is very poorly matched with any other term. From that we can deduce that the words that are present in texts that were labelled with spam are much different from the rest. This can be confirmed by the top keywords that were extracted from spam texts using the TF-IDF method. As we've already discussed in the earlier approach their content includes promotions and is not hostile opposed to the most other categories.

Then we've looked for how many terms is the vector most similar to the vector of the data category of the same name. That happened only to the term *racist*, therefore we've widen our approach to search for the top three terms. Then that was true for terms *racist, profane, harassment, insult* and *hostile*. For these terms we could have made a case that are more distinct from the rest and should be treated separately. On contrary, terms *offensive, abusive, cyberbullying, harrasment, obscene* and *discredit* were all most similar to the categories *hostile* and *derailing*. Therefore we've could presume that these terms are closely connected and are describing similar things.

Discussion & future directions

We observed that some categories (e.g. *profane, benevolent sexism, spam*) label documents, which truly have a distinct vocabulary when compared to other categories. But it is exactly the vocabulary and nothing more, which we compared using these sparse and non-contextual word embeddings. The fact is that documents from categories such as *hostile, offensive, discredit, hateful, insult* use quite similar vocabulary, but the context in which they are used, and the semantic interpretations can be quite different between these categories –

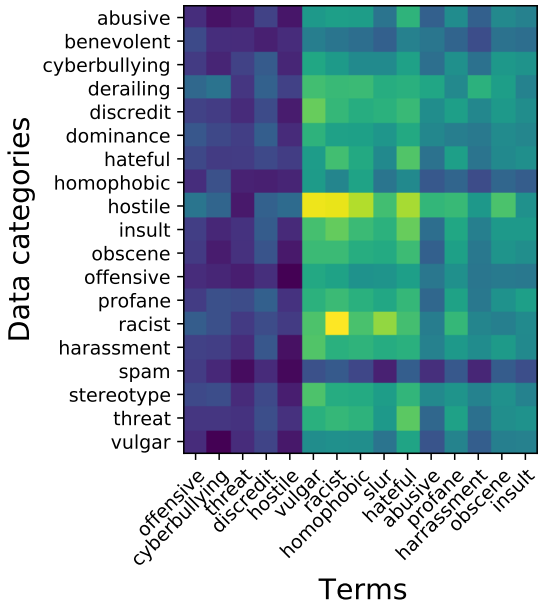

Figure 6. Similarity matrix of term Word2vec vectors and averaged Word2vec vectors obtained from categories. Lighter color marks higher similarity between two vectors. Similarity values range from 0.05 to 0.6.

this is why the documents were labeled differently in the first place. In order to explore such semantic differences between these categories, we will resort to contextual embeddings of sentences and documents, which can be obtained using e.g. BERT and ELMo. This, alongside multilingual embeddings will be our main research domain in the future.

References

[1] Sylvia Jaki and Tom De Smedt. Right-wing german hate speech on twitter: Analysis and automatic detection, 2019.

[2] A. Rodríguez, C. Argueta, and Y. Chen. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174, 2019.

[3] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings, 2019.

[4] Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!, 2020.

[5] Shervin Malmasi and Marcos Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.

[6] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, and Adebayo Abayomi-Alli. A probabilistic clustering model for hate speech classification in twitter. *Expert Systems with Applications*, page 114762, 2021.

[7] 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.

[8] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 2, pages 241–244, 2011.

[9] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press, 2018.

[10] 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*, FIRE '19, page 14–17, New York, NY, USA, 2019. Association for Computing Machinery.

[11] 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, 2018.

[12] Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data. [Accessed: 20. 03. 2020].

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

[14] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications, 2019.

[15] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, 2017.