Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6786–6794
Marseille, 11–16 May 2020

Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets

Paula Fortuna[1], **Juan Soler-Company**[1], **Leo Wanner**[2,1]
[1]NLP Group, Pompeu Fabra University, [2]Catalan Institute for Research and Advanced Studies (ICREA)
paulatfortuna@gmail.com
juan.soler|leo.wanner@upf.edu

Abstract

The field of the automatic detection of hate speech and related concepts has raised a lot of interest in the last years. Different datasets were annotated and classified by means of applying different machine learning algorithms. However, few efforts were done in order to clarify the applied categories and homogenize different datasets. Our study takes up this demand. We analyze six different publicly available datasets in this field with respect to their similarity and compatibility. We conduct two different experiments. First, we try to make the datasets compatible and represent the dataset classes as Fast Text word vectors analyzing the similarity between different classes in a intra and inter dataset manner. Second, we submit the chosen datasets to the Perspective API Toxicity classifier, achieving different performances depending on the categories and datasets. One of the main conclusions of these experiments is that many different definitions are being used for equivalent concepts, which makes most of the publicly available datasets incompatible. Grounded in our analysis, we provide guidelines for future dataset collection and annotation.

Keywords: hate speech, toxicity, aggression, offensive, dataset comparison

1. Introduction

Over the past years, the amount of online offensive speech has been growing steadily. To address this problem, a significant number of scientific publications focused on two different (albeit related) tasks: (i) the compilation and annotation of corpora and (ii) the automatic detection of different types of offensive speech, among them, e.g., toxicity, hate, abuse, using generic state-of-the-art (i.e., machine learning-based) natural language processing techniques. However, critics on this duality are starting to emerge in the community. The main concern resides in the restricted generalization potential of the trained machine learning (ML) models for classification of offensive speech. For instance, Swamy et al. (2019) find that by training on top of BERT models (Devlin et al., 2018), it is possible to obtain a language model that performs very competitively for different datasets, however, the generalization of the model depends highly on the training data. The authors hypothesize that a model will generalize better if it is used on data that is more similar to the data used for training, and worse if the data contains more non-offensive samples.

In general, the limited generalization potential is a problem that can have two different origins. Firstly, online offensive speech varies depending on the targeted groups (e.g., the terms used to express hate against a community in Africa are clearly different from the terms used against African Americans) and on the context (e.g., the hate against the LGBT community is likely to be worded differently in the context of a pro-Trump discussion and in the context of the Pride Parade). Secondly, ML-based techniques require large amounts of high quality annotated data – which raises the question about the annotation guidelines, and thus the clear definition of the categories to be annotated, as well as the compatibility of the categories across different datasets. According to vid (2019), the lack of clear definitions of key categories is a critical issue. The authors argue that researchers use different, sometimes theoretically ambiguous or misleading terms for equivalent categories. Thus, 'abusive' has been defined based on the speakers' intention to harm, which cannot always be determined by just looking at the content. Furthermore, definitions also make assumptions on the effect of the messages on the reader, which, obviously, depends entirely on the personality of the reader. The authors conclude that accurately defining key terms will result in better communication and collaboration in the field. Kumar et al. (2018) also point out that there is a large amount of terminology as well as different understandings of this terminology in the context of abusive speech. The fact that there are so many different definitions and interpretations of the same terms results in duplicated research, lack of clear goals and difficulties in reusing the data. The authors stress that it is of utmost importance that a common understanding of the problem is achieved such that standard datasets and different compatible approaches to solve the problem are developed. In another recent study (Swamy et al., 2019), it is also highlighted that more work must be done to identify similarities and differences in the publicly available datasets, as the data is more important than the model when tackling the problem of model generalization.

Our study takes up this demand. We analyze six different publicly available datasets on offensive speech in English, annotated in terms of a varying number of categories (including, e.g., 'hate speech', 'toxicity', 'sexism'), with respect to their similarity and compatibility and compare the performance of a state-of-the-art classification algorithm, which can be used via the Perspective API (Jigsaw, 2019a), on the different categories of these datasets. The outcome of our study signals that the intra-and inter-dataset coherence of the annotation should be improved. For this purpose, we provide guidelines for future dataset collection and annotation. The developed methodology for systematic comparison of the categories from different datasets can be also applied for validation of the labeling schemes of hate speech related language resources in general.

In what follows, we first briefly describe in Section 2. the setup of our study: the datasets we use, the methodology we apply, and the experiments. In Section 3., we discuss the outcome of the experiments. Section 4. presents the guidelines that are grounded in the assessment of the outcome of our experiments for future offensive dataset collection and annotation, and Section 5., draws some conclusions from our study and outlines some directions for future research.

2. Experimental Setup

2.1. Datasets

For this study, we use six publicly available datasets that cover different hate speech-related categories. Henceforth, the datasets are referred to as follows: *Waseem* (Waseem and Hovy, 2016), *Davidson* (Davidson et al., 2017), *Amievalita* (Fersini et al., 2018), *Hateval* (Basile et al., 2019), *TRAC* (Kumar et al., 2018), and *Toxkaggle* (Jigsaw, 2019b).

The main characteristics of the six datasets are shown in Table 1. Regarding the proportion of negative examples,[1] i.e., samples that do not belong to any of the targeted categories, to the proportions of samples that represent one of the targeted categories, in the case of the Waseem dataset, we observe that the majority of the data does not contain hate speech (68.02%) and the classes 'racism' and 'sexism' overlap, i.e., a number of messages are classified as both racist and sexist. In the Davidson dataset, the majority of the messages are either offensive or hateful; the percentage of the neutral messages ('neither') is very low (16.79%) when compared to the other datasets. According to the annotation schema of this dataset, offensive and hate speech are mutually exclusive.

The Amievalita dataset is more balanced: it contains 55.36% of non-misogynous messages. According to the Amievalita annotation scheme, misogyny is a super class of different subtypes of misogynistic behavior; the most common of them is the discredit (of women).

In the Hateval dataset, the majority of the messages contain no hate speech (57.97%). In this case, the class 'aggression' refers to a particular type of aggression within the hate speech context. This type of aggression is different than the one identified in the TRAC dataset, in which the majority of the messages contain some type of aggression, while only 42.10% contain no aggression.

For the Toxkaggle dataset, 10.16% of the messages contain some type of negative behavior. According to the annotation instructions for this dataset, 'severe toxicity', 'obscene', 'threat', 'insult' and 'identity attack' are subtypes of 'toxicity' (cf. Table 2). However, we noticed that the data is not consistent in this respect, as there are messages belonging to 'obscene' (N=317), 'insult' (N=301), 'identity hate' (N=54) and 'threat' (N=22), but not to 'toxicity'.

2.1.1. Class definitions

Table 2 shows the definitions of the individual categories as provided in the annotation guidelines for each dataset. The definitions of hate speech in the Waseem dataset, misogyny in the Amievalita dataset, and hate speech, misogyny and aggression in the Hateval dataset are explicit and precise since they aim to enumerate all possible cases that should be considered for the annotation of a given message in terms of a given category. Such explicitness is instrumental for high quality annotation. In contrast, the definitions of hate speech and offensive speech in the Davidson dataset are more vague. This has already been criticized by vid (2019), who pointed out that the term 'offensiveness' makes assumptions about the sensibility of the audience, which is intrinsically subjective. It implies the question: 'Offensive for whom?'. What is considered offensive by one audience, or in one context, might not be offensive elsewhere.

The TRAC definitions of overt and covert aggression are also very generic; covert aggression is simply defined as negation of overt aggression, which does not provide enough information about the class.

For the toxicity dataset provided in Toxkaggle, we could not find any specific definition of the categories. We assume that they are the same as the ones used in the context of the Perspective API as the developing team is the same (cf. Section 2.4.).

Another aspect to take into account is that it is often difficult to comprehend the difference between the labels 'aggression', 'toxicity' and 'offense'. They seem to be often used to refer to a general perception of pejorative speech. Similarly, it is difficult to grasp the difference between 'sexism' and 'misogyny'. Thus, in (Anzovino et al., 2018) misogyny is defined as "specific case of hate speech whose targets are women", which is very similar to the definition of sexist hate speech.

2.2. Label standardization between datasets

To be able to compare the different categories across the datasets and the results of our experiments that are described below, we standardize the label categories, assigning to the equivalent categories in the different datasets the same labels. The standardization takes into account the category definitions and the observations on these definitions in Subsection 2.1. Table 3 shows the standardization that was performed.

In the case of the Waseem dataset, the sexism and racism categories are considered to be subcategories of the hate speech category. Furthermore, the sexism category in this dataset is assumed to be equivalent to the misogynous category of the Amievalita dataset since in the literature no clear distinction between these two categories is provided. The resulting standardized cross-dataset label is called 'misoginy-sexism'. For the Davidson dataset, we created a new category 'toxicity' that subsumes the union of its hate speech and offensive categories.[2] Regarding the Hateval dataset, its aggression category covers a specific type of aggression as it is a subset of hate speech. In this

[1] For detailed frequencies see https://drive.google.com/open?id=1trR4rleMgL40YXwZgdxuzMExLwHBftOY

[2] Alternative labels would have been 'aggression' or 'abuse', which all are used in the literature for representing general classes.

Dataset id	Mutually Exclusive Classes	Classes	Data Collection Strategy	Number of Instances	Source	Reference
Waseem	no	racism, sexism	Initial search based on common slurs and terms used pertaining to religious, sexual, gender, and ethnic minorities.	16.914	Twitter	(Waseem and Hovy, 2016)
Davidson	yes	hate speech, offensive	Begining with the hatebase lexicon.	24.802	Twitter	(Davidson et al., 2017)
Amievalita	partially	misogynous, discredit, sexual harassment, stereotype, dominance, derailing	Search with representative slurs, monitoring of potential victims' and perpetrators' accounts.	4.000	Twitter	(Fersini et al., 2018)
Hateval	no	hate speech, aggression	Against immigrants and women.	9.000	Twitter	(Basile et al., 2019)
TRAC	yes	covert aggression, overt aggression	Searching for keywords and constructions that are often included in offensive messages, such as "she is", "antifa", "conservatives".	12.000	Facebook	(Zampieri et al., 2019b)
Toxkaggle	no	threat, identity hate, severe toxic, insult, obscene, toxic	Not provided	159.571	Wikipedia	(Jigsaw, 2019b)

Table 1: Dataset properties

dataset id	definitions
Waseem	"**Hate speech**: 1. uses a sexist or racial slur. 2. attacks a minority. 3. seeks to silence a minority. 4. criticizes a minority (without a well founded argument). 5. promotes, but does not directly use, hate speech or violent crime. 6. criticizes a minority and uses a straw man argument. 7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims. 8. shows support of problematic hash tags. E.g. #BanIslam, #whoriental, #whitegenocide 9. negatively stereotypes a minority. 10. defends xenophobia or sexism. 11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria." (Waseem and Hovy, 2016)
Davidson	"**Hate speech** is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. In extreme cases this may also be language that threatens or incites violence, but limiting our definition only to such cases would exclude a large proportion of hate speech. Importantly, our definition does not include all instances of **offensive language** because people often use terms that are highly offensive to certain groups but in a qualitatively different manner." ... "Such language is prevalent on social media (Wang et al. 2014), making this boundary condition crucial for any usable hate speech detection system." (Davidson et al., 2017)
Amievalita	Subtypes of **misogyny**: 1. **Stereotype & Objectification** is "a widely held but fixed and oversimplified image or idea of a woman; description of women's physical appeal andor comparisons to narrow standards." 2. **Dominance** is "to assert the superiority of men over women to highlight gender inequality." 3. **Derailing** is "to justify woman abuse, rejecting male responsibility; an attempt to disrupt the conversation in order to redirect women's conversations on something more comfortable for men." 4. **Sexual Harassment & Threats of Violence** is "to describe actions as sexual advances, requests for sexual favours, harassment of a sexual nature; intent to physically assert power over women through threats of violence." 5. **Discredit** is "slurring over women with no other larger intention." (Fersini et al., 2018)
Hateval	"**Hate Speech (HS)** is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. More specifically, HS against immigrants may include: 1. insults, threats, denigrating or hateful expressions. 2. incitement to hatred, violence or violation of rights to individuals or groups perceived as different for somatic traits (e.g. skin color), origin, cultural traits, language, etc. 3. presumed association of origin/ethnicity with cognitive abilities, propensity to crime, laziness or other vices. 4. references to the alleged inferiority (or superiority) of some ethnic groups with respect to others. 5. delegitimation of social position or credibility based on origin/ethnicity. 6. references to certain backgrounds/ethnicities as a threat to the national security or welfare or as competitors in the distribution of government resources. 7. dehumanization or association with animals or entities considered inferior. 8. the presence of aggressive language: the second one is on whether the tweet is aggressive or not. A message is considered **aggressive**, if: 1. it implies or legitimates discriminating attitudes or policies against the given target (immigrants/migrants/refugees). 2. there is an allusion to a potential threat posed by the presence of the target, or its alleged outnumbering with respect to the native population. 3. there is a sense of dissatisfaction and frustration, which may also result in overt hostility, due to the (perceived) privileged treatment granted to the target group by the government. 4. there is the reference (whether explicit or just implied) to violent actions of any kind perpetrated against the given target of the message. **Misogynous**: a text that expresses hating towards women in particular (in the form of insulting, sexual harassment, threats of violence, stereotype, objectification and negation of male responsibility). **Not Misogynous**: a text that does not express hating towards women in particular. IMPORTANT(!): a tweet is misogynous only if it is related to woman/women. **Aggressive**: a message is considered aggressive if it (implicitly or explicitly) presents, incites, threatens, suggests or alludes to: 1. attitudes, violent actions, hostility or commission of offenses against women; 2. justify or legitimize an aggressive action against women. **Not Aggressive**: If none of the previous conditions hold. (Basile et al., 2019)
TRAC	"Behaviours such as trolling, cyberbullying, flaming, insults, abusive / offensive language, hate speech, radicalization or racism have been analysed individually." ... "As we try to classify actual data in one of these categories, the overlap becomes even more prominent. As such it might be possible to tackle all of these using similar methods" ... **Overt aggression** is any speech / text (henceforth, text will mean both speech as well as text) in which aggression is overtly expressed - either through the use of specific kind of lexical items or lexical features which is considered aggressive and / or certain syntactic structures is overt aggression. **Covert aggression** is any text in which aggression is not overtly expressed is covert aggression. It is an indirect attack against the victim and is often packaged as (insincere) polite expressions (through the use of conventionalised polite structures), In general, lot of cases of satire, rhetorical questions, etc. may be classified as covert aggression." (Kumar et al., 2018)
Toxkaggle	Instructions in Kaggle: "You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of **toxicity** are: toxic, severe toxic, obscene, threat, insult, identity hate. You must create a model which predicts a probability of each type of toxicity for each comment.". (Jigsaw, 2019b)

Table 2: Conceptual definitions from the datasets analysed in this work.

case we do not merge both categories into 'toxicity' as this dataset aims to classify only hate speech, and considers aggression only when it happens in the context of hate speech.

The TRAC dataset contains the categories 'overt aggression' and 'covert aggression', which we merge into a new category 'aggression'. We could also convert it to a general category such as 'toxicity', however we opt not to do it as TRAC aims to identify subtler aggression, which is a dimension not mentioned in the Toxkaggle dataset.

2.3. Experiment 1: Analyzing categories

In the first experiment of this study, we aim to compare the categories across the annotated datasets with respect to both

dataset	original category	standardized category
Waseem	sexism	misoginy-sexism
	racism	racism
	sexism or racism	hate speech
Davidson	hate speech	hate speech
	offensive	offensive
	hate speech or offensive	toxicity
Amievalita	misogynous	misoginy-sexism
Hateval	hate speech	hate speech
	aggression	aggressive hate speech
TRAC	overt aggression	overt aggression
	covert aggression	covert aggression
	overt or covert aggression	aggression
Toxkaggle	insult	insult
	severe toxic	severe toxic
	obscene	obscene
	identity hate	hate speech
	threat	threat
	toxic	toxicity

Table 3: Category standardization.

their similarity to the other categories and their homogeneity, i.e., variation of the samples of one single category.

Each category is represented as a centroid vector using Fast Text (Bojanowski et al., 2016) and pretrained word embeddings trained on wikipedia (Mikolov et al., 2018). We decided to follow this approach because the majority of the datasets contain short texts generated on social networks and Fast Text along with pretrained word embeddings has been providing good results in diverse works applied to the automatic detection of hate speech and related concepts with similar data; see, e.g., (Santucci et al., 2018; Fortuna and Nunes, 2019).

The process that we use to compute the aforementioned centroids is as follows:

- Pre-process the messages by lowercasing all words, removing IPs, Twitter elements such as hashtags, usernames, and stop words using NLTK.
- Train word embeddings using FastText and the 300-dimension English wikipedia pretrained embeddings.
- Extract the centroid of the message by averaging the word embeddings of each of its sentences.

2.3.1. Inter-dataset class similarity

In this experiment, we aim to compare the different categories across the annotated datasets in terms of their semantic similarity. For this, in addition to the previous procedure we:

- Compute the average of every message centroid that belongs to each category, obtaining the centroid of each category.

After obtaining the category centroids, we perform a Principal Component Analysis (PCA) (Pearson, 1901) to obtain a 2D representation and thus be able to plot the centroids. The result of this process can be seen in Figure 1.

To complement the visualization of the category centroids, we also compute the distances between each pair of categories with the standardized category labels (see Table 3) to get a better grasp on how similar these categories actually are. To compute the distances, we use the cosine distance metric.

The analysis of the PCA plot and the inter-class distance analysis are presented in Subsections 3.1. and 3.2.. Both analyses are distinct and complementary. On one hand, the PCA represents the distance between classes when considering the feature reduction to two orthogonal dimensions. On the other side, the inter-class distance compares all messages of the corresponding classes. In other words, inter-class distance is a metric that measures the similarity between two classes in terms of how much the messages of the two vary.

2.3.2. Intra-dataset class homogeneity

In this experiment, we aim to compare the different categories across the annotated datasets with respect to their internal homogeneity. For this purpose, in addition to the procedure described in Subsection 2.3. we:

- Compute the distance between all the messages from the same category by using the cosine similarity.
- We then average all the distances in order to estimate the homogeneity of a category.

The analysis of the intra-class distances is presented in Subsection 3.3..

2.4. Experiment 2: Classifying with Perspective API

In order to analyse the generalization potential of a state-of-the-art model over the considered datasets and their categories, we use *Perspective API*. Perspective API was created by Jigsaw and Google's Counter Abuse Technology team in the context of the Conversation-AI project. The API provides several classifiers that compute scores between 0 and 1 for different categories (among others, 'toxicity' (Jigsaw, 2019a)), given an input text. The classifier uses Convolutional Neural Networks (CNNs) trained with GloVe word embeddings (Pennington et al., 2014) fine-tuned during training on data from online sources such as Wikipedia and The New York Times.

Perspective API provides the following definitions of the relevant categories:

- **toxicity** is a "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."
- **severe toxicity** is a "very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective."
- **identity attack** are "negative or hateful comments targeting someone because of their identity."
- **insult** is an "insulting, inflammatory, or negative comment towards a person or a group of people."

- **profanity** are “swear words, curse words, or other obscene or profane language”
- **threat** “describes an intention to inflict pain, injury, or violence against an individual or group.”

The Conversation-AI team at OffensEval-2019 applied Perspective API as a baseline system for toxicity detection, without any additional training on the contest data and obtained a very competitive result (12th out of 103 submissions, F1 of 0.79). This result encouraged us to choose this system to classify the six chosen datasets using the standardized categories and analyze its results.

For each standardized category in the datasets,[3] we evaluate how well the classifier is able to identify it and distinguish it from non-harmful messages. In other words, we perform binary classification using only the messages belonging to the analyzed category and the messages marked in the dataset as non-toxic, aggressive or in any way abusive. For the evaluation of the performance of the classifier on each dataset, we use the F1 metric.

3. Discussion

3.1. Centroid Visualization

To create the graph shown in Figure 1, we selected the two first principal components of the category centroids. The goal was to see how the different categories relate to each other, hence we plotted with a different color possibly related categories. The results seem to be coherent with what was expected as there are clear similarities between classes that represent similar categories. For instance, the ‘aggression’ related categories (in red) tend to be grouped together and intersected with hate speech, as expected for the ‘hateval-aggression’ category. The hate speech categories (in yellow) also appear close in space. From these categories, ‘davidson-hate speech’ and ‘hateval-hate speech’ are the closest, while ‘waseem-hate speech’ is at same time close to ‘hate speech’ but also between ‘waseem-sexism’ and ‘waseem-racism’ – again as expected.

The ‘toxkaggle-identity hate’ category appears close to ‘davidson-hate speech’, but, in this case, closer to other toxkaggle categories (‘toxic’, ‘insult’ and ‘obscene’). This is probably due to the multiclass property of the toxkaggle dataset, where the same message can have different labels. We think that this property, and also the fact that this is the only dataset collected from Wikipedia comments may justify why the toxicity dataset categories are mapped together in the upper part of the figure and are more difficult to compare to the categories of the other datasets. However, at same time ‘toxkaggle-severe-toxic’ is far from the same dataset categories, including the ‘toxkaggle-toxic’.

Apart from conforming an expected degree of similarity between specific categories of the different datasets, the PCA plot allows us to gather new insights about the data. For instance, the general categories ‘toxicity’ from toxkaggle (‘toxkaggle-toxic’) and ‘aggression’ from TRAC (‘trac-CAG’ or ‘trac-OAG’) do not appear close, despite the fact that both toxicity and aggression are defined as general umbrella terms for offensive, toxic or abusive online behavior. In contrast, ‘toxkaggle-toxic’ and ‘davidson-toxicity’, which in our category standardization were assigned the label ‘toxicity’, appear closer in the plot. Additionally, between these two categories, ‘amievalita-sexism-misoginy’ is situated, indicating that ‘sexism’ can be one of the main types of toxicity in those datasets.

Also, the ‘misoginy-sexism’ related categories (‘amievalita-sexism-misoginy’ and ‘waseem-sexism’) seem close, but ‘davidson-offensive’ categories seem more similar to those. Another interesting observation is that the category ‘waseem-racism’ seems to be very close to both TRAC dataset categories, indicating that racism can be more represented than other categories in the TRAC dataset .

3.2. Inter Dataset Class Distance

To further analyze how similar or dissimilar the categories across the datasets are, let us look at the distances between class centroids. Table 4 shows each category of each dataset (in bold as header) and the top 5 most similar categories (below the header).[4] As expected, these results are aligned with the PCA. For the hate speech related categories, we see that ‘Davidson-hate speech’ is close to ‘toxkaggle-identity hate’ and Hateval’s ‘hate speech’, but farther from Waseem’s ‘hate speech’, ‘racism’, ‘sexism’ and ‘Amievalita-misogyny’. This seems to indicate that there are several different representations of the notions of ‘hate speech’ and its subtypes.

‘Amievalita-misogynous’ appears to be close to Davidson’s ‘offensive’ and Toxkaggle’s ‘toxicity’, but it is also not so far away from Waseem’s ‘sexism’. On the other side, ‘waseem-sexism’ is also closer to ‘toxkaggle-toxic’ than to ‘amievalita-misogynous’. This may indicate that the Toxkaggle ‘toxicity’ category contains sexist messages that are more similar to the ‘waseem-sexist’ messages. Nevertheless, it is unexpected that Waseem’s ‘sexism’ category appears more similar to ‘toxicity’ than to ‘amievalita-misogyny’. Some further analysis would be needed to understand why.

Regarding the Toxkaggle categories, its ‘identity hate’ is close to its ‘insult’, ‘toxic’ and ‘obscene’, and more distant to the hate speech categories from the other datasets (i.e., ‘davidson-hate speech’ and ‘hateval-hate speech’, or ‘amievalita-sexism-misogyny’). This indicates that in this dataset the category notions are very interdependent. Even more obvious is the overlap between Toxkaggle’s ‘insult’ and ‘obscene’, which are very close to each other and largely share the distances to the other categories. Indeed, the distinction between both is not clear.

Toxkaggle’s ‘severe toxic’ is closer to all the other Toxkaggle’s dataset categories, but the reverse does not apply. Thus, ‘toxkaggle-toxic’ is closer to ‘toxkaggle-insult’, followed by ‘toxkaggle-identity hate’ and ‘amievalita-misogyny’, and very far from ‘severe toxic’, which is quite

[3] Due to the API quota limits, we randomly sampled 20% of the Toxkaggle dataset in a total of 31.914 messages.

[4] The full table with the distance values for each pair of categories can be found at https://docs.google.com/spreadsheets/d/1mkSTmuO8cc8tUbAEq68J_el39hyx6uvEWo1xPFGMRvg/edit?usp=sharing.

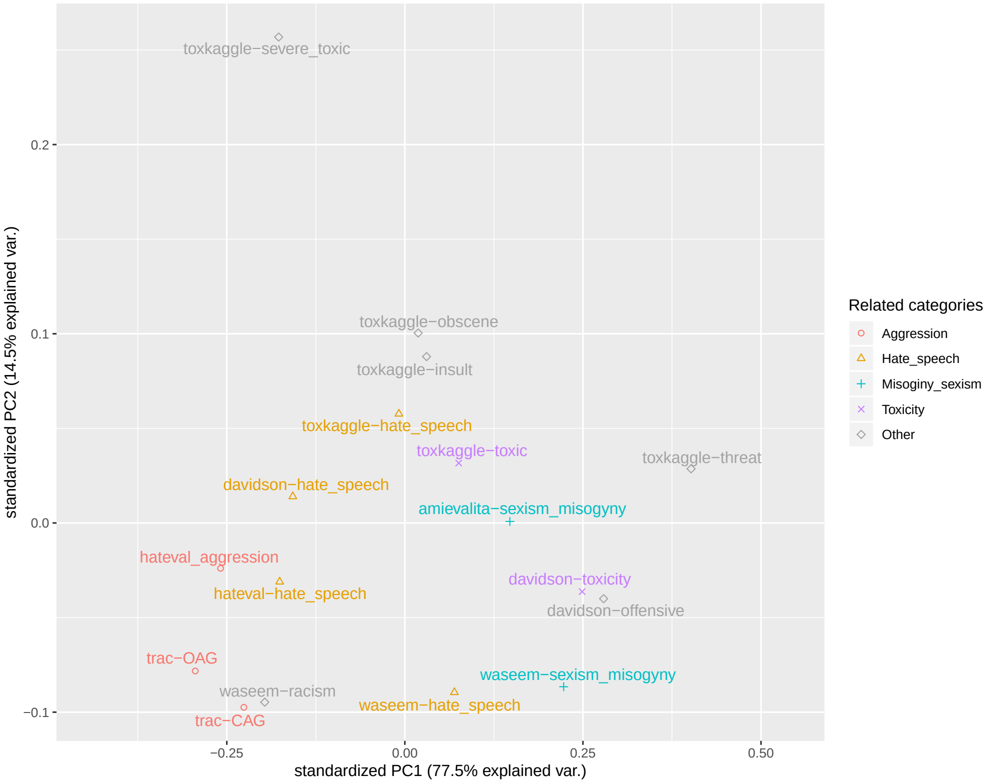

Figure 1: PCA results

unexpected, since their labels suggest that the main difference between these two categories is the intensity of the expressed toxicity.

3.3. Intra-Category Homogeneity

Figure 2 displays the homogeneity of the individual categories for each dataset in terms of the average distance between their messages. The more homogeneous a category is, the smaller the value in the plot.

We can observe that the most homogeneous category is ‘waseem-racism’. Specific types of harmful content such as racist, misogynous and threats appear to be quite homogeneous, which indicates that these categories are well-defined and its messages are clearly identifiable. On the other hand, hate speech presents various homogeneity scores: ‘waseem-hate_speech’ is quite homogeneous, since it is composed of racist and misogynous messages while Davidsons’ hate speech instances are very heterogeneous, which is coherent with its definition, where messages that express hatred towards any target group are considered (see Table 2).

The assessment of the number of messages per category shows that the homogeneity is not affected by it. Furthermore, homogeneity does not depend on the dataset, neither on the platform used for data collection.

3.4. Results Perspective API

Figure 3 shows the results of the classification experiment with Perspective API. The results reveal that the performance of the classifier has a huge variation depending on the category. The classifier is better at identifying ‘toxicity’, ‘offense’, followed by ‘obscene’, ‘insult’, ‘misogyny-sexism’, ‘hate speech’, and worse at identifying ‘aggression’, ‘racism’, ‘severe toxic’ and ‘threat’.

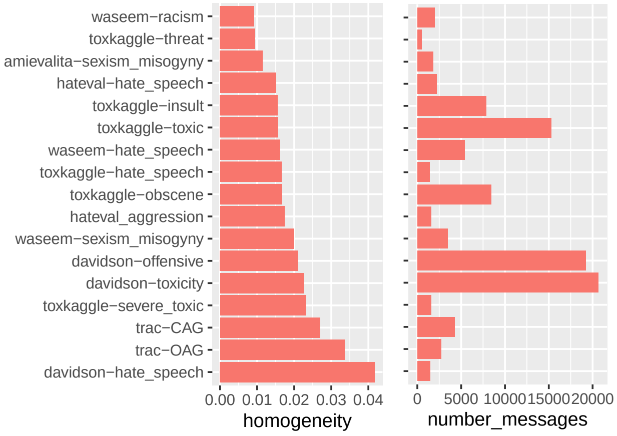

Figure 2: Class homogeneity and number of messages per class.

It is interesting to notice that despite the fact that the Perspective API classifiers draw upon the same categories as used in the Toxkaggle dataset for annotation, the classifier seems to handle better ‘misoginy-sexism’ than ‘racism’.

Additionally, we can notice that for the same category, the performance has high variability across datasets. For instance, for the hate speech category, the classifier shows a higher F1 for the Davidson dataset than for the Hateval dataset. The performance is even worse for the case of the

trac-cag	**trac-oag**	**davidson-toxicity**	**davidson-hate_speech**
trac-oag	trac-cag	davidson-offensive	toxkaggle-identity_hate
waseem-racism	hateval-aggression	amievalita-misogynous	hateval-hate_speech
hateval-aggression	waseem-racism	waseem-sexism	hateval-aggression
hateval-hate_speech	hateval-hate_speech	waseem_hate_speech	toxkaggle-toxic
waseem_hate_speech	waseem_hate_speech	toxkaggle-toxic	toxkaggle-obscene
davidson-offensive	**amievalita-misogynous**	**hateval-aggression**	**hateval-hate_speech**
davidson-toxicity	davidson-toxicity	hateval-hate_speech	hateval-aggression
amievalita-misogynous	davidson-offensive	waseem-racism	waseem-racism
waseem-sexism	toxkaggle-toxic	trac-oag	trac-cag
waseem_hate_speech	waseem-sexism	trac-cag	waseem_hate_speech
toxkaggle-toxic	toxkaggle-insult	davidson-hate_speech	trac-oag
toxkaggle-identity_hate	**toxkaggle-insult**	**toxkaggle-obscene**	**toxkaggle-threat**
toxkaggle-insult	toxkaggle-obscene	toxkaggle-insult	toxkaggle-toxic
toxkaggle-toxic	toxkaggle-toxic	toxkaggle-toxic	amievalita-misogynous
toxkaggle-obscene	toxkaggle-identity_hate	toxkaggle-identity_hate	toxkaggle-insult
davidson-hate_speech	amievalita-misogynous	amievalita-misogynous	waseem-sexism
hateval-hate_speech	hateval-hate_speech	hateval-hate_speech	toxkaggle-obscene
toxkaggle-severe_toxic	**toxkaggle-toxic**	**waseem_hate_speech**	**waseem-racism**
toxkaggle-obscene	toxkaggle-insult	waseem-sexism	hateval-aggression
toxkaggle-insult	toxkaggle-identity_hate	hateval-hate_speech	hateval-hate_speech
toxkaggle-identity_hate	toxkaggle-obscene	toxkaggle-toxic	trac-cag
toxkaggle-toxic	amievalita-misogynous	waseem-racism	trac-oag
davidson-hate_speech	waseem_hate_speech	amievalita-misogynous	waseem_hate_speech
waseem_sexism			
waseem_hate_speech			
toxkaggle-toxic			
amievalita-misogynous			
davidson-toxicity			
davidson-offensive			

Table 4: Top 5 most similar to each label

Toxkaggle dataset. This confirms that each dataset provides its own flavor of hate speech.

For the general categories, ‘toxicity’ and ‘aggression’, the classifier achieves a higher F1 on the Davidson dataset than on the Toxkaggle dataset. The performance is even worse for the TRAC ‘aggression’. This means that, indeed, the ‘aggression’ category as used in the TRAC dataset cannot be compared and merged with the ‘toxicity’ category.

Also, when we compare the performance on the Toxkaggle dataset, we can see that it performs better when applied to categories with more instances in the dataset such as ‘toxic’, ‘obscene’ and ‘insult’ and worse when applied to smaller categories such as ‘hate speech’, ‘severe toxic’ and ‘threat’. This indicates that the sampling procedure has a direct impact on the performance of the classifier, as better-represented classes are clearly better identified.

Our experiment also confirms Kumar et al. (2018)’s observation that covert aggression (‘trac-CAG’) is recognized worse than overt aggression (‘trac-OAG’).

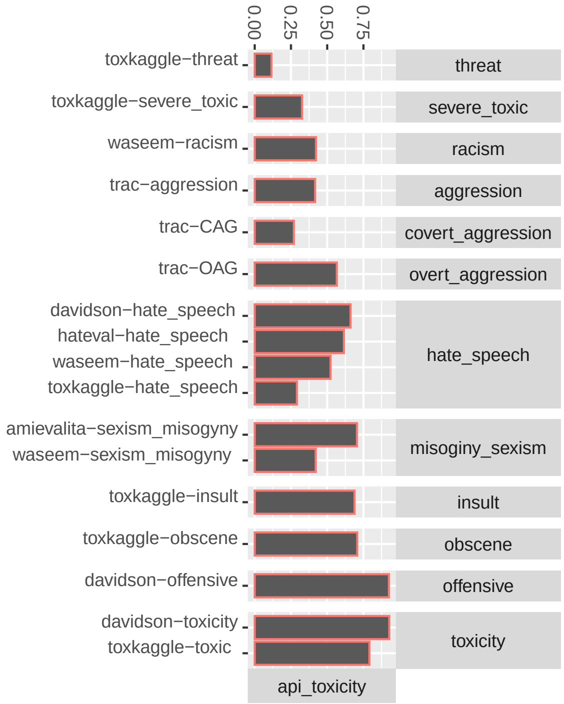

Figure 3: Toxicity’s Perspective API classification performance by category (F1 metric).

4. Guidelines for Future Concept Definition and Dataset Annotation

The literature review and the qualitative analysis conducted in this study show that dataset quality in the field of hate speech should be improved. The results of our analysis also suggest that the intra and inter-dataset coherence of the annotation should equally be improved. The following guidelines are intended as the first step to address both problems:

- Strive for clear and distinctive category definitions. That is, definitions should be more similar to the definition of hate speech in the Waseem dataset (Waseem and Hovy, 2016), misogyny in the Amievalita dataset (Fersini et al., 2018) and hate speech, misogyny and aggression in the Hateval (Basile et al., 2019) than to definitions in (Davidson et al., 2017; Jigsaw, 2019a)).
- Avoid creating new categories to refer to concepts already present in the literature. In the case a new category is identified, provide clear examples and justification why a new category is needed.
- Position new categories in the map of existing categories when annotating new datasets, for instance, by following a similar method to the one provided in this

work.

- Develop hierarchical multiclass annotation schemas, as already called for in (Fortuna et al., 2019; Zampieri et al., 2019a; Jigsaw, 2019b). Multiclass schemas facilitate the development of targeted classifiers for the different types of pejorative online behavior. Generic classifiers, trained on binary annotation schemas, do not address the coverage of fine-grained categories (such as, e.g., 'threat').

- Collect information that allows to control possible dataset bias, such as, e.g., the profile of the author of the message. Cf. (Arango et al., 2019), where it is shown that hate speech related datasets are collected from a limited set of authors, such that when messages from the same author are only in the training set, the performance of the models drops, showing the influence of the author's writing style on the model. Controlling the aforementioned bias via the introduction of features to counterbalance it will improve the model generalization.

- Provide detailed information on the sampling procedure. For instance, the used source (e.g., Twitter), the groups targeted by the authors (e.g., women), the context the data refers to (e.g., comments on news about politics, or sports), the time and location for the data.

- Provide detailed information on the class balancing procedure. We saw that the proportion between offense, toxicity, abuse or hate messages can vary across different datasets, and this factor greatly impacts the classifiers' performance.

5. Conclusions

The presented study aimed at comparing a selection of publicly available datasets and clarifying the categories of these datasets – a question that has been already raised in the literature. This study has been framed as an analysis of category definitions and running two distinct experiments. The outcome of the first experiment, was that hate speech related categories seem to be more coherent and similar between themselves and the same for aggression related categories. However, the categories aiming at the representation of classes that cover all types of pejorative online speech, such as toxicity and aggression, do not seem related among each other. The second experiment, which implied the use of the Perspective API classifier, showed that even when datasets use very generic categories, their diverging definitions, data samples or inconsistent annotation may lead to diverging classifier performance – as, e.g., in the case of 'aggression' from the TRAC dataset and 'toxicity' from the Toxkaggle dataset. Our theoretical and empirical analysis gave rise to guidelines that we hope will help to bring some more clarity in the context of the annotation of pejorative online speech. However, an even deeper exploration of the data in the future would be helpful. For instance, it would be worth to explore why the 'waseem-racism' category is very close in the PCA plot to the TRAC 'aggression' categories, or why 'waseem-sexism' is, according to this plot, more similar to 'toxicity' than to 'misogyny', or what makes most of the studied categories rather heterogenous (in contrast to, e.g., âwaseem-racismâ). One possible approach would be to use different methods for feature extraction and see how different categories would map in relation to the other.

Acknowledgements: We thank the reviewers for their insightful comments. The first author is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program *Human Capital* (POCH), supported by the European Social Fund and by national funds from MCTES. The work of the second and third authors has been supported by the European Commission in the context of the H2020 Research Program under the contract numbers 700024 and 786731.

References

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 45–54. ACM.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@ CLiC-it*.

Fortuna, P. Soler-Company, J. and Nunes, S. (2019). Stop propaghate at semeval-2019 tasks 5 and 6: Are abusive language classification results reproducible? In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Jigsaw. (2019a). Perspective api. Available in `https://github.com/conversationai/`

perspectiveapi, accessed last time in November 2019.

Jigsaw. (2019b). Toxic comment classification challenge: Identify and classify toxic online comments. Available in https://www.kaggle.com/c/jigsaw -toxic-comment-classification-challenge, accessed last time in November 2019.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Santucci, V., Spina, S., Milani, A., Biondi, G., and Di Bari, G. (2018). Detecting hate speech for italian language in social media. In *EVALITA 2018, co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China, November. Association for Computational Linguistics.

(2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.