Resources and Techniques for User and Author Profiling in Abusive Language (ResT-UP 2020), pages 14–19
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Lower Bias, Higher Density Abusive Language Datasets: A Recipe

Juliet van Rosendaal, Tommaso Caselli, Malvina Nissim
University of Groningen
Oude Kijk in't Jaatstraat, 26 9712 EK Groningen
julietlucienne@msn.com, {t.caselli,m.nissim}@rug.nl

Abstract

Datasets to train models for abusive language detection are both necessary and scarce. One reason for their limited availability is the cost of their creation. Manual annotation is expensive, and on top of it, the phenomenon itself is sparse, causing human annotators having to go through a large number of irrelevant examples in order to obtain some significant data. Strategies used until now to increase density of abusive language and obtain more meaningful data, include data filtering on the basis of pre-selected keywords and hate-rich sources of data. We suggest a recipe that at the same time can provide meaningful data with possibly higher density of abusive language and also reduce top-down biases imposed by corpus creators in the selection of the data to annotate. More specifically, we exploit the controversy channel on Reddit to obtain keywords that are used to filter a Twitter dataset. While the method needs further validation and refinement, our preliminary experiments show a higher density of abusive tweets in the filtered *vs.* unfiltered datasets, and a more meaningful topic distribution after filtering.

1. Problem Statement

The automatic detection of abusive and offensive messages in on-line communities has become a pressing issue. The promise of Social Media to create a more open and connected world is challenged by the growth of abusive behaviors, among which cyberbullying, trolling, and hate speech are some of the most known. It has also been shown that awareness of being a victim of some kind of abusive behavior is less widespread than what one actually reports as having experienced (Jurgens et al., 2019).

The body of work conducted in the areas of abusive language, hate speech, and offensive language has rapidly grown in the last years, leaving the field with a variety of definitions and a lack of reflection on the intersection among such different phenomena (Waseem et al., 2017; Vidgen et al., 2019). As a direct consequence, there has been a flood of annotated datasets in different languages, [1] all somehow addressing the same phenomena (e.g. offensive language, or hate speech) but applying slightly different definitions, different annotation approaches (e.g. experts *vs.* crowdsourcing), and different reference domains (e.g., Twitter, Facebook, Reddit). Hate speech, in particular, has been the target of the latest major evaluation campaigns such as SemEval 2019 (Zampieri et al., 2019b; Basile et al., 2019), EVALITA 2018 (Bosco et al., 2018), and IberEVAL 2018 (Fersini et al., 2018) in an attempt to promote both the development of working systems and a better understanding of the phenomenon.

Vidgen et al. (2019) and Jurgens et al. (2019) identify a set of pending issues that require attention and care by people in NLP working on this topic. One of them concerns a revision of what actually constitutes abuse. The perspective that has been adopted so far in the definition of abusive language, and most importantly of hate speech, has been limited to specific and narrow types of abusive/hateful behaviors to recognize. For instance, definitions of hate speech have been carefully carved, focusing on the intentions of the message producer and by listing cases of applications (e.g., attack against an individual or a group on the basis of race, religion, ethnic origin, sexual orientation, disability, or gender). As a consequence, more subtle but still debasing and harmful cases are excluded, and (potential) negative effects of the messages on the targets are neither considered nor accounted for.

A further problematic aspect in previous work concerns the quality of the datasets. Besides issues on the annotation efforts (i.e., amount of data and selected annotation approach), one outstanding problem is the collection of data. While some language phenomena are widespread in any (social media) text one may collect (e.g. presence of named entities), hate speech is not. Random sampling from targeted platforms is thus a non-viable solution as it will entail going through a large amount of non-hateful messages before finding, very sparse, hateful cases. To circumvent this obstacle, three main strategies have been adopted so far:

- use of communities (Tulkens et al., 2016; Merenda et al., 2018): potentially hateful or abusive messages are extracted by collecting data from on-line communities that are known either to promote or tolerate such types of messages;
- use of keywords (Waseem and Hovy, 2016; Basile et al., 2019; Zampieri et al., 2019a): specific keywords which are not hateful or abusive *per se* but that may be the target of hateful or abusive messages, like for instance the word "migrants", are selected to collect random messages from Social Media outlets;
- use of users (Wiegand et al., 2018; Ribeiro et al., 2018): seed users that have been identified via some heuristics to regularly post abusive or hateful materials are selected and their messages collected. In a variation of this approach, additional potential "hateful" users are identified by applying network analysis to the seed users.

[1] For a more detailed overview of available datasets in different languages please consult `https://github.com/leondz/hatespeechdata`.

Common advantages of these approaches mainly lie in the reduction of annotation time and a higher density of positive instances, i.e. hateful messages in our case. However, a common and non-negligible downside is the developer's bias that unavoidably seeps in the datasets, although with varying levels of impact. For instance, it has been shown that Waseem and Hovy (2016) is a particularly skewed datasets with respect to topics and authors (Wiegand et al., 2019). For instance, words such as "commentator", "comedian", or "football" have strong correlations with hateful messages, or that hateful messages are mainly distributed across 3 different authors.

In this contribution, we present a simple data-driven method towards the creation of a corpus for hate speech annotation. We apply it to Dutch, a less resourced language for this phenomenon, but the method can be conceived as a blueprint to be applied to any other language for which social media data are available.

Our approach exploits cross-information from Twitter and Reddit, mainly relying on tf-idf and keyword matching. Through a series of progressive refinements, we show the benefits of our approach through a simple qualitative analysis. Finally, results of a trial annotation experiment provide further support for the proposed method.

Contributions We summarise our contributions as follows:

1. a bottom-up approach to collect potential abusive and hateful messages on Twitter by using keywords based on controversial topics emerging from a different social media platform, Reddit, rather than manually selected by developers;

2. promote the cross-fertilisation of different language domains (i.e., Twitter and Reddit), facilitate the identification of implicit forms of abusive language or hate speech, and reduce top-down bias by avoiding preselection of keywords by dataset creators;

3. work towards the development of a reference corpus for Dutch annotated for abusive language and hate speech.

2. A Possible Solution

Finding instances of abusive or hateful messages in Social Media is not an easy task. Founta et al. (2018) has estimated that abusive messages represent between 0.1% and 3% (at most) of the messages in Twitter. Furthermore, one of our goals is to propose a methodology to improve the collection of potentially abusive messages across Social Media platforms, independently from their specific characteristics. For instance, the community-based approach can be easily applied on Social Media such as Facebook or Reddit since Facebook pages and sub-reddits can be interpreted as proxies for communities of users that share the same interests. However, such an approach cannot be applied on Twitter where such an aggregation of users is not possible given the peculiar structure of the platform.

Previous work (Graumans et al., 2019), however, has shown that controversies can actually be used as a viable proxy to collect and aggregate abusive language from Social Media, especially Twitter. Indeed, controversies are interactions among individuals or groups where the opinions of the involved parties do not change and tend to become more and more polarised towards extreme values (Timmermans et al., 2017). Such a dynamic of interactions and their polarised nature is a potential growth medium for abusive language and hate speech. A further advantage of using controversies to collect data is the reduction of topic bias factors. Although the proposed method will still use keywords to identify the data, such keywords have not been manually selected by the developers of the datasets but they are learned in a bottom-up approach from data that are perceived by the public at large or Social Media communities as divisive and potentially subject to a more extreme style of expression.

We focus on Twitter data rather than other Social Media platforms for a number of reasons, among which the most relevant are: (1.) possibility of (re-)distributing the data to the public, in compliance with the platform's terms of use and EU GDPR regulations; (2.) popularity of the platform in previous work on abusive language and hate speech, thus facilitating comparisons across languages and the development of cross-lingual models; (3.) ease of access to the data.

2.1. Method Overview

We conducted two initial experiments that could allow the identification of controversial topics on Twitter and thus extract potential abusive and hateful messages. The unfiltered Twitter dataset contains all public Dutch tweets posted in August 2018, corresponding to 14,122,350 tweets.

Twitter-based hashtag filter As an initial exploratory experiment, we tested whether using the N most frequent hashtags over a period of time could be a viable solution. The working hypothesis being: the more frequent the hashtag, the more likely it may refer to a controversy. We set the time frame to 1 month (i.e., August 2018), identified the most frequent hashtags (not necessarily corresponding to the trending topics in the targeted time span) and collected all tweets that contained them. The approach was quite a failure, as we mainly extracted tweets generated by bots and by account of professional institutions (e.g. news outlets), rather than actual users. We immediately dismissed this approach.

Reddit-based bag-of-words filter. This second experiment adopts a more refined approach and contextually investigates *cross-information* of Social Media platforms. We turned our attention on Reddit, a social media platform organised around specific channels ('subreddits'), using its filtering tools. Reddit allows its users to upvote and downvote posts, which resolves in a democratic procedure to give topics that deserve more attention precedence over topics considered less important. The tools can filter on top posts, thus showing the posts with the most upvotes, as well as on the so-called "controversial" posts, showing posts with a more or less equal amount of upvotes and downvotes. This is basically showing that the opinions on the relevance of the posts are mixed. We then retrieved two datasets: one of which was filtered on top posts (**top**), and another which was filtered on controversial posts

(**controversial**), with no time restriction (i.e. use of the "all time" option). The top dataset contains 48 posts (for a total of 279,057 words) while the controversial dataset, contains 20 submissions (with a total of 23,794 words). All posts were taken from `r/thenetherlands`, a subreddit with 237,000 subscribers at the time of this study and with mainly Dutch contributions.

We then extracted unigram keywords per dataset using TF-IDF. In particular, we calculated TF-IDF over the union of the two datasets, i.e., **top** ∪ **controversial**, then we selected the k most important unigrams relative to each dataset, and retained only those of the controversial one. This procedure represents the core aspect of our bottom-up approach to select relevant keywords for highly controversial topics.

We then applied the controversial keywords to filter the 14M Twitter dataset extracting all messages that contain at least one of them. Next to this procedure, we also implemented a secondary filter based on the hashtags of all the extracted messages. We applied these additional set of hashtag-based keywords to retrieve additional messages from the 14M Twitter dataset. A visualization of the process is shown in Figure 1.

The final amount of collected messages by applying the two sets of keywords is 784,000 tweets (corresponding to 5.6% of the original 14M messages). A manual exploration of a portion of the new dataset has shown that the messages were actually referring to controversial topics and their origin was mainly from actual users rather than bots or by accounts of institutions.

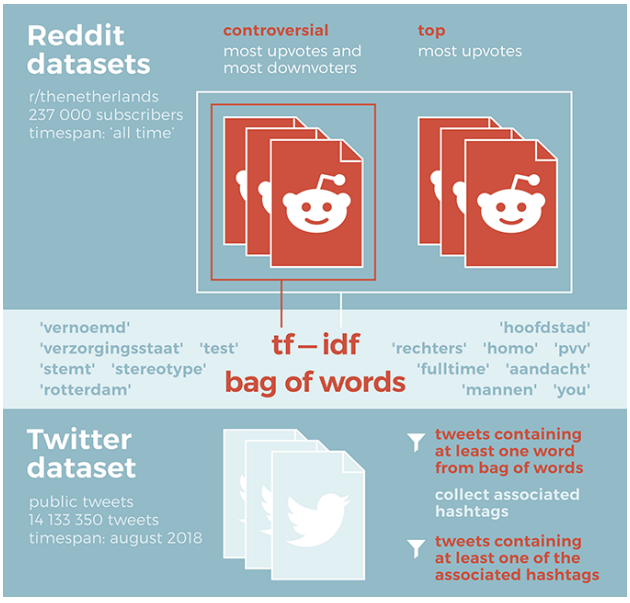

Figure 1: Reddit-based filtering process

3. Validation

After concluding that our second attempt seemed promising enough, we conducted a validation step to verify whether the filtering renders a higher density of tweets with abusive or hate speech instances. In addition, we also wanted to verify whether the filtered dataset potentially contained more interesting tweets for the abusive language and hate speech detection tasks. For the density aspect, we conducted a double annotation over a small random selection of 500 tweets from the filtered dataset and 500 tweets from the unfiltered one (Section 3.1.). For the qualitative aspect, we simply created word clouds of the two different sets of tweets, and observed which token would stand out most (Section 3.2.). This would give a rough but immediate idea of the most present topics in the two sets.

3.1. Annotation

We annotated the data by using a simplified version of the guidelines for hate speech annotation developed by Sanguinetti et al. (2018). We only considered the annotation parameter of hate speech [yes/no]. A tweet that would be annotated as containing hate speech should have a clear target of a minority group and should be "spreading, inciting, promoting or justifying hatred or violence toward the target, or aiming at dehumanizing, delegitimating, hurting or intimidating the target", as taken from the guidelines of (Sanguinetti et al., 2018).

To give some examples of tweets from the filtered dataset that were perceived as challenging to annotate:

1. Iedere scholier die toch een telefoon bij zich heeft/gebruikt op school krijgt 10 zweepslagen en meer bij recidivering. Maar dat zal wel niet mogen van die slappe homo's van @groenlinks @d66 hollandsezaken
   Every student carrying/using a mobile phone at school receives 10 whiplashes or more in case of recurrence. But the whimpy fags from @groenlinks @d66 probably won't allow that. hollandsezaken

2. RT @hulswood: Moskee-organisatie NL neemt Turkse jongeren mee op trainingkamp radicale imam: "trouw met zesjarig kind, mannen mogen vrouwen slaan, en steun gewapende jihad Syrië". ....was te verwachten, dit is islam. NL moet islamisering actief stoppen!
   RT @hulswood: Dutch mosque organization takes Turkish youth to training camp of radical imam: "marry a six year old, men are allowed to beat women, support the armed jihad in Syria". ... this was to be expected, this is Islam. The Netherlands has to actively stop islamization!

3. Schandalig om een hond met deze hitte aan een boom vast te binden. Doe je toch ook met pvv'ers niet?
   It is scandalous to tie a dog to a tree in this heat. You woudn't do that with a politician from the PVV either, right?

Though still low, a higher proportion of hate speech tweets was found in the filtered dataset. In Table 1 we show the confusion matrix for the two annotators over the two sets. After discussion and reconciliation, the total number of hateful tweets was 7 for the unfiltered dataset and 18 for the filtered one. There is a margin of disagreement that suggests further annotation is necessary, and for the moment led to interesting findings, also regarding the annotation guidelines.

Figure 2: Word cloud of unfiltered dataset (125 words are shown)

Figure 3: Word cloud of filtered dataset (125 words are shown)

Non filtered dataset		
	a1: 'no'	a1: 'yes'
a2: 'no'	491	1
a2: 'yes'	7	1

Filtered dataset		
	a1: 'no'	a1: 'yes'
a2: 'no'	464	4
a2: 'yes'	24	8

Table 1: Annotation confusion matrices for both datasets (before discussion and reconciliation).

The discussion over disagreements between the annotators showed an extra parameter that could possibly be taken into account (next to target and action) for the annotation guidelines, namely goal, that can be seen both as writer's intentions and message's effect on receivers. One annotator pointed out how for certain tweets no actual hate speech was expressed, e.g. the action of "spreading, inciting, promoting or justifying hatred or violence toward the target, or aiming at dehumanizing, delegitimating, hurting or intimidating", though the intentions of the user and the effects of the message could be interpreted as doing so. On the other hand, the other annotator had marked such tweets as non hate speech.

To clarify this issue consider the following example:

4. RT @SamvanRooy1: Qua symboliek kan dit tellen: in het Nederlandse Deventer verdwijnt een synagoge door toedoen van de gemeente en een Turkse ondernemer. Moslims erin, Joden eruit: bij gelijkblijvend beleid is dat het West-Europa van de toekomst. Video. islamisering
RT @SamvanRooy1: Symbolically this could count: a synagogue is taken out of service in the Dutch city

Deventer, because of the municipality and a Turkish businessman. Muslims in, Jews out: if this policy remains is this what West-Europe of the future looks like. Video. islamization

As Twitter is already using a hate speech filter, the tweets that are easier to track down are possibly already filtered out. For example, tweets with curses or death threats were not found. Tweets with less explicit, but more suggestive or subtle abusive language is left. Whether or not one can go as far to proclaim these to be hate speech is a challenging judgement, which could benefit from more elaborate and/or precise annotation guidelines. For instance, one useful distinction could be to annotate the explicitness of messages against a target rather than having a binary hate speech distinction (Waseem et al., 2017).

3.2. Topics

In Figure 2 and in Figure 3 we show the word clouds for the unfiltered and filtered datasets, respectively (125 words each). Any comment we can make about the two clouds is simply qualitative and should require a more structured analysis and further annotation.

At first sight, we can observe that in the filtered set, several of the words can indeed be signalling controversial topics. Examples are political parties (pvv, d66), politicians such as Wilders (wilders) (Dutch far-right politicians) and Rutte (rutte) (prime minister), morokkan (Moroccan), islam (Islam), feministen. The unfiltered set does not lend itself equally easily to meaningful clusters, showing quite generic, neutral terms such as echt (true) and genoeg (enough). Another quite clear example of this contrast between more specific *vs*. more generic in the two sets is provided by 'people' terms: the unfiltered set shows mensen ('people') and kinderen ('children'), while in the filtered set we find quite dominantly the terms for 'men' (mannen) and 'women' (vrouwen).

Some other terms can possibly be interpreted in connection with the time the Tweets were collected (August 2018), but with some degree of speculation. During that period, Amsterdam hosted the gay pride, which could have been the object of controversial comments. Rotterdam could be connected to the Rotterdam Rave Festival. Both sets show a reference to politie (police) that would require further analysis for proper understanding.

4. Future Directions

The recipe we have proposed here to maximise annotation effort over a meaningful and denser dataset for detecting abusive language, and to contextually minimise data selection bias, is only in its first experimental tests. However, we believe our results are promising and deserve further investigation, especially since this methodology could be applied to any language for which one can obtain Twitter and (controversial) Reddit data.

First, we need to annotate more data to confirm that the filtered dataset has indeed both a higher concentration of abusive language as well as overall a more interesting semantic profile, which ensures a more focused and challenging task. This need is also prompted by some discrepancy between the annotators; this is standardly observed in hate speech annotation, but we need to better understand whether filtering (or not) affects disagreement, and in which way. Second, we want to further explore and understand the potential of cross-fertilisation between different social media platform. This would also imply singling out and assessing the actual contribution of this aspect within our proposed recipe. Would it also be possible to use yet other platforms? Could we induce the filtering keywords through other channels maintaining our bottom-up strategy? Lastly, but importantly, we need to assess the actual quality of the filtered *vs*. unfiltered datasets in terms of training data for abusive language detection. Are we indeed creating 'better' data for predictive models? For a proper test of this sort, the test data would need to be acquired independently of our suggested strategy, which however could incur the classic problem of top-down bias which we wanted to avoid in the first place. This test clearly requires proper modelling, possibly under different settings.

5. Acknolwedgement

We would like to thank Arjan Schelhaas for contributing to the annotation, and the anonymous reviewers for their helpful comments.

6. Bibliographical References

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., 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, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Bosco, C., Dell'Orletta, Felice, F. P., Sanguinetti, M., and Tesconi, M. (2018). Overview of the EVALITA Hate Speech Detection (HaSpeeDe) Task. In Tommaso Caselli, et al., editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Fersini, E., Rosso, P., and Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Graumans, L., David, R., and Caselli, T. (2019). Twitter-based polarised embeddings for abusive language detection. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 9.

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy, July. Association for Computational Linguistics.

Merenda, F., Zaghi, C., Caselli, T., and Nissim, M. (2018). Source-driven Representations for Hate Speech Detection, proceedings of the 5th italian conference on computational linguistics (clic-it 2018). Turin, Italy.

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., and Meira Jr, W. (2018). Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Timmermans, B., Aroyo, L., Tobias Kuhn, Kaspar Beelen, E. K., and Bob van de Velde, G. v. E. (2017). Controcurator: Understanding controversy using collective intelligence. In *Collective Intelligence 2017*.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016). A dictionary-based approach to racism detection in dutch social media. In *Proceedings of the first Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)/Daelemans, Walter [edit.]; et al.*, pages 1–7.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August. 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.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018). Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, 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. In *Proceedings of NAACL*.

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). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.