HateBERT

Caselli, Tommaso; Basile, Valerio; Mitrović, Jelena; Granitzer, Michael

Published in:
ArXiv

Publication date:
2020

HateBERT: Retraining BERT for Abusive Language Detection in English

Tommaso Caselli[♣], **Valerio Basile**[◇], **Jelena Mitrović**[‡], **Michael Granitzer**[‡]
[♣]University of Groningen, [◇]University of Turin, [‡]University of Passau
Groningen The Netherlands, Turin Italy, Passau Germany
◇{valerio.basile}@unito.it, ♣t.caselli@rug.nl
‡jelena.mitrovic|michael.granitzer}@uni-passau.de

Abstract

In this paper, we introduce HateBERT, a re-trained BERT model for abusive language detection in English. The model was trained on RAL-E, a large-scale dataset of Reddit comments in English from communities banned for being offensive, abusive, or hateful that we have collected and made available to the public. We present the results of a detailed comparison between a general pre-trained language model and the abuse-inclined version obtained by retraining with posts from the banned communities on three English datasets for offensive, abusive language and hate speech detection tasks. In all datasets, HateBERT outperforms the corresponding general BERT model. We also discuss a battery of experiments comparing the portability of the general pre-trained language model and its corresponding abusive language-inclined counterpart across the datasets, indicating that portability is affected by compatibility of the annotated phenomena.

1 Introduction

The widespread popularity of social media and micro-blogging platforms is still having undisclosed effects in our life as a result of an increased connectivity among people. However, the potential benefits are overshadowed by numerous expressions of offensive and abusive language. This contribution focuses on finding solutions to overcome that problem.

Hate speech, offensive and abusive language have recently become topics of widespread interest in the Natural Language Processing (NLP) community, as shown by the development of datasets in multiple languages (Waseem and Hovy, 2016; Poletto et al., 2017; Founta et al., 2018; Zampieri et al., 2019a; Ibrohim and Budi, 2019; Sigurbergsson and Derczynski, 2020; Çöltekin, 2020; Pitenis et al., 2020), dedicated workshops[1] and evaluation campaigns (Wiegand et al., 2018; Bosco et al., 2018; Zampieri et al., 2019b; Basile et al., 2019). This interest has resulted in a fragmented picture of the various language phenomena at stake accompanied by a variety of definitions and (in)compatibility of the annotations (Waseem et al., 2017).

The development of systems for the automatic identification of abusive and offensive language have followed a common trend in NLP: feature-based linear classifiers (Waseem and Hovy, 2016; Ribeiro et al., 2018), neural network architectures (e.g., CNN or Bi-LSTM) (Kshirsagar et al., 2018; Mishra et al., 2018; Mitrović et al., 2019), and, finally, fine-tuning pre-trained language models, e.g., BERT, RoBERTa, among others (Liu et al., 2019; Swamy et al., 2019). Results vary both across datasets and architectures, with linear classifiers qualifying as very competitive, if not better, when compared to neural networks. On the other hand, systems based on pre-trained language models have proven to have the best performance in this area, reaching new state-of-the-art results. One issue with these pre-trained models is that the training language variety makes them well suited for general-purpose language understanding tasks. To address this, there is a growing interest in generating domain-specific BERT-like pre-trained language models, such as AlBERTo (Polignano et al., 2019), a Twitter-based BERT model in Italian, BioBERT for the biomedical domain in English (Lee et al., 2019), or FinBERT for the financial domain in English (Yang et al., 2020). In this work, we introduce HateBERT, a pre-trained BERT model for investigating hate speech, offensive and abusive language in social media.

[1]The Workshop on Online Abuse and Harms - fourth edition https://www.workshopononlineabuse.com/home

Another relevant aspect is the generalisability of trained models, typically hindered by the lack of harmonization among datasets and annotation schemes. Previous work (Karan and Šnajder, 2018; Benk, 2019; Pamungkas and Patti, 2019; Rizoiu et al., 2019) has addressed this task by conflating generalisability with portability. Datasets with different phenomena have been forced into homogenous annotations by collapsing different labels into (binary) macro-categories. At the same time, different methods have been applied, including data augmentation on the line of (Daumé III, 2007) or the integration of transfer learning techniques. In this paper, we present the results of a set of experiments across datasets, showing how differences in the annotated phenomenon affect models' portability while saying little about generalisability.

In summary, the main contributions of this work are the following:

- a large-scale dataset of social media posts in English from communities banned for being offensive, abusive, or hateful;
- a comparison between fine-tuned systems based on a general pre-trained language model and a abusive language-inclined version;
- a battery of experiments showing that differences in the annotated phenomena affects portability of models and say little about their generalisability;

2 HateBERT: Re-training BERT with Abusive Online Communities

Transformer-based pre-trained language models, such as BERT, are the most recent wave of state-of-the-art architectures applied to address NLP tasks. While generally achieving good performance on numerous NLP tasks, when applied to less standard language varieties, such as social media data, results may fluctuate a lot. For instance, by comparing different fine-tuned BERT models on the OffensEval 2019 dataset (Zampieri et al., 2019b), it appears that the key factor in boosting the performance is the quality of the pre-processing step (Liu et al., 2019; Swamy et al., 2019), rather than other aspects such as learning rate or training time.

Given the limited size of task-specific datasets that can be used to fine-tune these architectures, it appears that retraining such models to shift their representations towards occupying a space closer both to the language variety and to the targeted phenomenon is a viable and cheaper solution, rather than manually annotating more data.

Previous work in this direction (Polignano et al., 2019; Lee et al., 2019; Yang et al., 2020) has used massive amounts of data that could be easily recovered (random tweets, PubMED articles, and financial documents respectively) either for training BERT-like language models from scratch or re-training an existing model. However, when it comes to hate speech or offensive and abusive language, options for suitable (i.e., large) and representative (i.e., language, domain and/or targeted phenomenon) datasets are limited. Directly scraping messages containing profanities is not the best option as lots of potentially useful data may be missed. Graumas et al. (2019) suggest scraping tweets about controversial topics to generate offensive-loaded embeddings, but their approach presents some limits since the offensive-loaded embeddings do not beat general embeddings in their experiments. On the other hand, Merenda et al. (2018) have shown the effectiveness of using messages from potentially hateful on-line communities to generate so-called "hate embeddings". We followed this latter approach by using messages from banned communities in Reddit to re-train a general BERT model.

RAL-E: the Reddit Abusive Language English dataset Reddit is one of the most popular social media where users share and discuss content. The website is organized into over one million user-created and user-moderated communities known as *subreddits*. In 2015, Reddit strenghtened its anti-harassment policy banning several subreddits on multiple occasions (Chandrasekharan et al., 2017). We retrieved a large list of banned communities in English from different sources including official posts by the Reddit administrators and dedicated Wikipedia pages.[2] We selected only communities that were banned because deemed to host or promote offensive and/or abusive content (e.g., expressing harassment, bullying, inciting/promoting violence, inciting/promoting hate). We collected the posts from the selected

[2] https://en.wikipedia.org/wiki/Controversial_Reddit_communities

communities by crawling a collection of Reddit Comments from December 2005 to March 2017.[3] For each post, we kept the text and its metadata, including the username of the author and the timestamp. The resulting collection comprises 6,955,084 messages from a period between 2012 and 2017. The complete list of selected communities and the number of messages retrieved per community is reported in Table 4 in the Appendix.

Creating HateBERT From the RAL-E dataset we obtained a collection of ≈71,1 million tokens. We then re-trained the general English BERT `base uncased` model[4] by applying the Masked Language Model (MLM) objective and using the default parameters. The result is a shifted BERT model, HateBERT `base uncased`, along two dimensions: (i.) language variety (i.e. social media); and (ii.) polarity (i.e., offensive- and abusive-oriented model).

3 Experiments and Results

To verify the validity of HateBERT as being more suitable than a general one, i.e. BERT, for detecting offensive and abusive language phenomena, we run a set of experiments on three English datasets.

OffensEval 2019 (Zampieri et al., 2019b) This dataset was distributed in the context of the SemEval 2019: Task 6[5] evaluation exercise. The dataset contains 14,100 tweets annotated for **offensive** language. The dataset is split into training and test, with 13,240 messages in training and 860 in test. The positive class (i.e. messages labeled as offensive) are 4,400 in training and 240 in test.

AbusEval (Caselli et al., 2020) This dataset has been obtained by adding a layer of **abusive** language annotation to OffensEval 2019. The overall size of the dataset is the same as OffensEval 2019, i.e., 14,100 tweets, as well as that of the training and test splits (13,240 and 860 messages, respectively). On the other hand, the differences concern the distribution of the positive class (i.e., messages labeled as abusive) which results in 2,749 in training and 178 in test.

HatEval (Basile et al., 2019) This dataset was distributed for the SemEval 2019: Task 5[6] evaluation exercise. The English portion of the dataset contains 13,000 tweets annotated for **hate speech** against migrants and women. The training set is composed of 10,000 messages while the test data contains 3,000 messages. Both training and test contain an equal amount of messages with respect to the targets, i.e., 5,000 each in training and 1,500 each in test. This does not hold for the distribution of the positive class (i.e., messages annotated as hateful with respect to the specific target) where 4,165 messages are present in the training and 1,252 in the test set.

A common characteristic of the datasets is the imbalance between positive and negative classes, as an attempt to provide a more realistic distribution of the targeted phenomena messages in the real world.[7] At the same time, these datasets provide annotations for three different phenomena. This allows us to evaluate the robusteness of the abusive language-inclined pre-trained language model, as well as to investigate the extent to which the portability of models is influenced by differences in annotations.

We used the same pre-processing steps and hyperparameter settings when fine-tuning each of the pre-trained models (BERT *vs*. HateBERT). Hyperparameters (Table 3) and pre-processing steps are more closely detailed in the Appendix. In the fine-tuning step, we added a linear classifier on top of the pooled output for the [CLS] token to generate the predictions.

Table 1 illustrates the results on each dataset (in-dataset evaluation), while Table 2 reports on the cross-dataset evaluations. All results are averaged over 5 runs.

The in-domain results confirm the validity of the re-training approach as a strategy to generate better models for offensive and abusive language detection. On all datasets, HateBERT largely outperforms the

[3] `https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/`

[4] We used the pre-trained model available via the huggingface Transformers library - `https://github.com/huggingface/transformers`

[5] `https://competitions.codalab.org/competitions/20011`

[6] `https://competitions.codalab.org/competitions/19935`

[7] The actual distribution of offensive/abusive/hateful messages in a platform like Twitter is estimated between 1% and 3% (Founta et al., 2018).

Dataset	Model	Macro F1	F1 (Negative class)	F1 (Positive class)
OffensEval 2019	BERT	.803	.892	.715
	HateBERT	**.805**	**.895**	.715
AbusEval	BERT	.724	.905	.542
	HateBERT	**.742**	**.910**	**.574**
HatEval	BERT	.480	.328	.633
	HateBERT	**.494**	**.352**	**.615**

Table 1: BERT *vs*. HateBERT: in-dataset evaluation. Best scores in bold.

Dataset	Model	OffensEval 2019	AbusEval	HatEval
OffensEval 2019	BERT	–	.726	.545
	HateBERT	.–	.732	.544
AbusEval	BERT	.710	–	.611
	HateBERT	.722	–	.619
HatEval	BERT	.572	.590	–
	HateBERT	.562	.582	–

Table 2: BERT *vs*. HateBERT: Cross-dataset evaluation (macro-F1) for all dataset combinations. Rows show the dataset used to train the model and columns the dataset used for testing. Best scores per training/test combination are underlined.

corresponding general BERT model. A detailed analysis of the results per class show that the improvements, in all datasets, affect both the positive and the negative classes, suggesting that HateBERT is more robust. Interestingly, the use of data from a different social media platform does not harm the fine-tuning stage of the retrained model, opening up possibilities of cross-fertilisation studies across social media platforms.

The cross-dataset results are less clear-cut, with BERT and HateBERT obtaining comparable results with a 50% split of the cases where one model outperforms the other. However, if we focus on the compatibility of the annotated phenomena, HateBERT appears to produce more portable models. For instance, fine-tuned models with HateBERT are more portable when using AbusEval, a dataset targeting a language phenomenon more specific than offensive language but less specific than hate speech.[8] Therefore, *the less compatible the annotated phenomena are (e.g. offensive language vs. hate speech), the better a general model works*. However, none of these results can be interpreted in the light of generalisability of a language phenomenon, because models are applied to datasets containing compatible but different phenomena.

4 Conclusion and Future Directions

This contribution introduces HateBERT `base uncased`,[9] an abusive language-inclined BERT for developing more robust systems for the automatic detection of hate speech, offensive and abusive language. The re-training step is done using the RAL-E dataset, a collection of ≈71,1 million tokens from banned communities in Reddit, by applying the Masked Language Model (MLM) objective. The in-dataset evaluation shows that HateBERT clearly outperforms the general BERT when fine-tuned on three different datasets, each representing a different language phenomenon, such as offensive language (OffensEval 2019), abusive language (AbusEval), and hate speech (HatEval). The cross-dataset experiments return a less clear picture of the behavior of HateBERT, suggesting that portability is influenced by compatibility of annotations and say little, if nothing, about generalisability (i.e., how good is the performance of a different data distribution annotated with the same guidelines and for the same phenomenon?).

Future work will focus on two directions: (i.) investigating to what extent the embedding representations of HateBERT are actually different from a general BERT pre-trained model, and (ii.) further testing the generalisability and portability of HateBERT fine-tuned models.

[8] Hate speech can be framed as a specific case of abusive language.

[9] HateBERT, the fine-tuned model, and the RAL-E dataset will be made publicly available.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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.

Michaela Benk. 2019. *Data Augmentation in Deep Learning for Hate Speech Detection in Lower Resource Settings*. Ph.D. thesis, Universität Zürich.

Cristina Bosco, Fabio Poletto Dell'Orletta, Felice, Manuela Sanuguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA Hate Speech Detection (HaSpeeDe) Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France, May. European Language Resources Association.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*. ELRA.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1:1–22, 12.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Leon Graumas, Roy David, and Tommaso Caselli. 2019. Twitter-based Polarised Embeddings for Abusive Language Detection. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–7.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium, October. Association for Computational Linguistics.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium, October. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. Source-driven Representations for Hate Speech Detection. In *Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10, Brussels, Belgium, October. Association for Computational Linguistics.

Jelena Mitrović, Bastian Birkeneder, and Michael Granitzer. 2019. nlpUP at SemEval-2019 task 6: A deep neural language model for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 722–726, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *CEUR WORKSHOP PROCEEDINGS*, volume 2006, pages 1–6. CEUR-WS.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. Hate speech detection through alberto italian language understanding model. In Mehwish Alam, Valerio Basile, Felice Dell'Orletta, Malvina Nissim, and Nicole Novielli, editors, *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AIIA 2019), Rende, Italy, November 19th-22nd, 2019*, volume 2521 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. 2019. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 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.

Zeerak Waseem and Dirk Hovy. 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.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. *Austrian Academy of Sciences, Vienna September 21, 2018*.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 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.

Appendix

Hyperparameter	Value
Learning rate	1e-5
Training Epoch	5
Adam epsilon	1e-8
Max sequence length	100
Batch size	32
Num. warmup steps	0

Table 3: Hyperparamters for fine-tuning BERT and HateBERT.

Pre-processing before fine-tuning For each dataset, we have adopted minimal pre-processing steps. In particular:

- all users' mentions have been substituted with a placeholder (@USER);
- all URLs have been substituted with a with a placeholder (URL);
- emojis have been replaced with text (e.g. 🥺 → `:pleading_face:`) using Python `emoji` package;
- hashtag symbol has been removed from hasthtags (e.g. #kadiricinadalet → kadiricinadalet);
- extra blank spaces have been replaced with a single space.

Subreddit	Number of posts	Subreddit	Number of posts
CringeAnarchy	3,627,030	N1GGERS	663
fatpeoplehate	1,585,112	FULLFASCISM	545
uncensorednews	617,954	Identitarians	529
milliondollarextreme	593,049	niggerspics	501
sjwhate	175,256	niglets	411
WhiteRights	126,289	Rapefugees	397
GreatApes	79,472	niggervideos	342
Physical_Removal	32,620	TheGoyimKnow	340
TrayvonMartin	23,077	WatchNiggersDie	311
europeannationalism	21,933	KKK	303
NationalSocialism	13,068	polacks	221
holocaust	11,441	niggas	192
nazi	7,609	NatSoc	190
pol	6,196	NiggerDrama	175
Truecels	5,757	NiggerFacts	172
Polistan	3,798	Chimpout	163
Braincels	3,650	USBlackCulture	115
GasTheKikes	3,469	niggersstories	87
RapingWomen	2,010	ChimpireOfftopic	51
ZOG	1,927	ShitNiggersSay	50
BlackCrime	1,359	Detoilet	47
misogyny	1,216	NiggersNews	46
GentilesUnited	1,163	BritishJewishPower	44
PhilosophyOfRape	1,116	chimpmusic	40
DylannRoofInnocent	1,092	funnyniggers	36
hitler	1,049	NiggersTIL	31
AganistGayMarriage	839	NiggerCartoons	31
niggerhistorymonth	28	Apefrica	26
Fuck_Niggers	26	teenapers	24
gibsmedat	24	didntdonuffins	24
WTFniggers	23	Homophobes	22
chicongo	22	NegroFree	20
blackpeoplehate	20	whitesarecriminals	20
TNB	18	muhdick	18
RacistNiggers	17	The_Nazi	17
NiggerMythology	15	NiggersGIFs	14
far_right	14	Quranimals	12
JustBlackGirlThings	11	TheRacistRedPill	11
QAnon	11	RacoonsAreNiggers	8
RapingEllenPao	7	NiggerDocumentaries	6
klukluxklan	6	niggerrebooted	6
apewrangling	5	TIL_4_Niggers	5
IHateWhitePeople	5	beatingfaggots	3
kike	2	RapeWorthy_Feminists	2
ChimpireMETA	2	BlackHusbands	1
NiggerSafari	1	ChimpinAintEasy	1
killniggers	1	AsianFemaleHate	1
killthejews	1		

Table 4: Distribution of messages per banned community composing the RAL-E dataset.