HateBERT: Retraining BERT for Abusive Language Detection in English

Tommaso Caselli [1], **Valerio Basile** [2], **Jelena Mitrović** [3], **Michael Granitzer** [3]
[1] University of Groningen [2] University of Turin [3] University of Passau
[1] t.caselli@rug.nl [2] valerio.basile@unito.it
[3] {jelena.mitrovic, michael.granitzer}@uni-passau.de

Abstract

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 curated and made available to the public. We present the results of a detailed comparison between a general pre-trained language model and the retrained version 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 fine-tuned models across the datasets, suggesting that portability is affected by compatibility of the annotated phenomena.

1 Introduction

The development of systems for the automatic identification of abusive language phenomena has followed a common trend in NLP: feature-based linear classifiers (Waseem and Hovy, 2016; Ribeiro et al., 2018; Ibrohim and Budi, 2019), neural network architectures (e.g., CNN or Bi-LSTM) (Kshirsagar et al., 2018; Mishra et al., 2018; Mitrović et al., 2019; Sigurbergsson and Derczynski, 2020), and fine-tuning pre-trained language models, e.g., BERT, RoBERTa, a.o., (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 reached 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, and it highlights their limits with more domain-specific language varieties. 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) or TweetEval (Barbieri et al., 2020) for Twitter, BioBERT for the biomedical domain in English (Lee et al., 2019), FinBERT for the financial domain in English (Yang et al., 2020), and LEGAL-BERT for the legal domain in English (Chalkidis et al., 2020). We introduce HateBERT, a pre-trained BERT model for abusive language phenomena in social media in English.

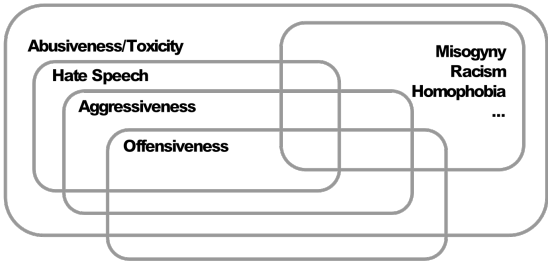

Figure 1: Abusive language phenomena and their relationships (adapted from Poletto et al. (2020)).

Abusive language phenomena fall along a wide spectrum including, a.o., microaggression, stereotyping, offense, abuse, hate speech, threats, and doxxing (Jurgens et al., 2019). Current approaches have focus on a limited range, namely offensive language, abusive language, and hate speech. The connections among these phenomena have only superficially been accounted for, resulting in a fragmented picture, with a variety of definitions, and (in)compatible annotations (Waseem et al., 2017). Poletto et al. (2020) introduce a graphical visualisation (Figure 1) of the connections among abusive language phenomena according to the definitions in previous work (Waseem and Hovy, 2016; Fortuna and Nunes, 2018; Malmasi and Zampieri, 2018; Basile et al., 2019; Zampieri et al., 2019). When it comes to offensive language, abusive language, and hate speech, the distinguishing factor is their level of specificity. This makes offensive language

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 17–25

the most generic form of abusive language phenomena and hate speech the most specific, with abusive language being somewhere in the middle. Such differences are a major issue for the study of portability of models. Previous work (Karan and Šnajder, 2018; Benk, 2019; Pamungkas and Patti, 2019; Rizoiu et al., 2019) has addressed this task by conflating portability with generalizability, forcing datasets with different phenomena into homogenous annotations by collapsing labels into (binary) macro-categories. In our portability experiments, we show that the behavior of HateBERT can be explained by accounting for these difference in specificity across the abusive language phenomena.

Our key contributions are: (i.) additional evidence that further pre-training is a viable strategy to obtain domain-specific or language variety-oriented models in a fast and cheap way; (ii.) the release of HateBERT, a pre-trained BERT for abusive language phenomena, intended to boost research in this area; (iii.) the release of a large-scale dataset of social media posts in English from communities banned for being offensive, abusive, or hateful.

2 HateBERT: Re-training BERT with Abusive Online Communities

Further pre-training of transformer based pre-trained language models is becoming more and more popular as a competitive, effective, and fast solution to adapt pre-trained language models to new language varieties or domains (Barbieri et al., 2020; Lee et al., 2019; Yang et al., 2020; Chalkidis et al., 2020), especially in cases where raw data are scarce to generate a BERT-like model from scratch (Gururangan et al., 2020). This is the case of abusive language phenomena. However, for these phenomena an additional predicament with respect to previous work is that the options for suitable and representative collections of data are very limited. Directly scraping messages containing profanities would not be the best option as lots of potentially useful data may be missed. Graumas et al. (2019) have used tweets about controversial topics to generate offensive-loaded embeddings, but their approach presents some limits. On the other hand, Merenda et al. (2018) have shown the effectiveness of using messages from potentially abusive-oriented on-line communities to generate so-called *hate embeddings*. More recently, Papakyriakopoulos et al. (2020) have shown that biased word embeddings can be beneficial. We follow the idea of exploiting biased embeddings by creating them using messages from banned communities in Reddit.

RAL-E: the Reddit Abusive Language English dataset Reddit is a popular social media outlet where users share and discuss content. The website is organized into user-created and user-moderated communities known as *subreddits*, being *de facto* on-line communities. In 2015, Reddit strengthened its content policies and banned several subreddits (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 Wikipedia pages.[1] We then selected only communities that were banned for being deemed to host or promote offensive, abusive, and/or hateful content (e.g., expressing harassment, bullying, inciting/promoting violence, inciting/promoting hate). We collected the posts from these communities by crawling a publicly available collection of Reddit comments.[2] For each post, we kept only the text and the name of the community. The resulting collection comprises 1,492,740 messages from a period between January 2012 and June 2015, for a total of 43,820,621 tokens. The vocabulary of RAL-E is composed of 342,377 types and the average post length is 32.25 tokens. We further check the presence of explicit signals of abusive language phenomena using a list of offensive words. We selected all words with an offensiveness scores equal or higher than 0.75 from Wiegand et al. (2018)'s dictionary. We found that explicit offensive terms represent 1.2% of the tokens and that only 260,815 messages contain at least one offensive term. RAL-E is skewed since not all communities have the same amount of messages. The list of selected communities with their respective number of retrieved messages is reported in Table A.1 and the top 10 offensive terms are illustrated in Table A.2 in Appendix A.

Creating HateBERT From the RAL-E dataset, we used 1,478,348 messages (for a total of 43,379,350 tokens) to re-train the English BERT `base-uncased` model[3] by applying the Masked Language Model (MLM) objective. The remaining 149,274 messages (441,271 tokens) have been used as test set. We retrained for 100 epochs (al-

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

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

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

most 2 million steps) in batches of 64 samples, including up to 512 sentencepiece tokens. We used Adam with learning rate `5e-5`. We trained using the huggingface code[4] on one Nvidia V100 GPU. 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., offense-, abuse-, and hate-oriented model).

Since our retraining does not change the vocabulary, we verified that HateBERT has shifted towards abusive language phenomena by using the MLM on five template sentences of the form "`[someone] is a(n)/ are [MASK]`". The template has been selected because it can trigger biases in the model's representations. We changed `[someone]` with any of the following tokens: "you", "she", "he", "women", "men" Although not exhaustive, HateBERT consistently present profanities or abusive terms as mask fillers, while this very rarely occurs with the generic BERT. Table 1 illustrates the results for "women".

BERT	HateBERT
"women"	
excluded (.075)	stu**d (.188)
encouraged (.032)	du*b (.128)
included (.027)	id***s (.075)

Table 1: MLM top 3 candidates for the templates "Women are [MASK.]".

3 Experiments and Results

To verify the usefulness of HateBERT for detecting abusive language phenomena, we run a set of experiments on three English datasets.

OffensEval 2019 (Zampieri et al., 2019) the dataset contains 14,100 tweets annotated for **offensive** language. According to the task definition, a message is labelled as offensive if "it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct." (Zampieri et al., 2019, pg. 76). 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. No development data is provided.

AbusEval (Caselli et al., 2020) This dataset has been obtained by adding a layer of **abusive** language annotation to OffensEval 2019. Abusive language is defined as a specific case of offensive language, namely "hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions." (Caselli et al., 2020, pg. 6197). The main difference with respect to offensive language is the exclusion of isolated profanities or untargeted messages from the positive class. The size of the dataset is the same as OffensEval 2019.The differences concern the distribution of the positive class which results in 2,749 in training and 178 in test.

HatEval (Basile et al., 2019) The English portion of the dataset contains 13,000 tweets annotated for **hate speech** against migrants and women. The authors define hate speech 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." (Basile et al., 2019, pg. 54). Only hateful messages targeting migrants and women belong to the positive class, leaving any other message (including offensive or abusive against other targets) to the negative class. The training set is composed of 10,000 messages and the test contains 3,000. 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, where 4,165 messages are present in the training and 1,252 in the test set.

All datasets are imbalanced between positive and negative classes and they target phenomena that vary along the specificity dimension. This allows us to evaluate both the robusteness and the portability of HateBERT.

We applied the same pre-processing steps and hyperparameters when fine-tuning both the generic BERT and HateBERT. Pre-processing steps and hyperparameters (Table A.3) are more closely detailed in the Appendix B. Table 2 illustrates the results on each dataset (in-dataset evaluation), while Table 3 reports on the portability experiments (cross-dataset evaluation). The same evaluation metric from the original tasks, or paper, is applied, i.e., macro-averaged F1 of the positive and negative classes.

The in-domain results confirm the validity of the re-training approach to generate better models for detection of abusive language phenomena, with HateBERT largely outperforming the corre-

[4] https://github.com/huggingface/transformers/tree/master/src/transformers

Dataset	Model	Macro F1	Pos. class - F1
OffensEval 2019	BERT	.803±.006	.715±.009
	HateBERT	**.809±.008**	**.723±.012**
	Best	.829	.752
AbusEval	BERT	.727±.008	.552±.012
	HateBERT	**.765±.006**	**.623±.010**
	Caselli et al. (2020)	.716±.034	.531
HatEval	BERT	.480±.008	.633±.002
	HateBERT	**.516±.007**	**.645±.001**
	Best	.651	.673

Table 2: BERT *vs.* HateBERT: in-dataset. Best scores in bold. For BERT and HateBERT we report the average from 5 runs and its standard deviations. *Best* corresponds to the best systems in the original shared tasks. Caselli et al. (2020) is the most recent result for AbusEval.

Train	Model	OffensEval 2019	AbusEval	HatEval
OffensEval 2019	BERT	–	.726	.545
	HateBERT	.–	.750	.547
AbusEval	BERT	.710	–	.611
	HateBERT	.713	–	.624
HatEval	BERT	.572	.590	–
	HateBERT	.543	.555	–

Table 3: BERT *vs.* HateBERT: Portability. Columns show the dataset used for testing. Best macro F1 per training/test combination are underlined.

sponding generic model. A detailed analysis per class shows that the improvements affect both the positive and the negative classes, suggesting that HateBERT is more robust. 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-fertilization studies across social media platforms. HateBERT beats the state-of-the-art for AbusEval, achieving competitive results on OffensEval and HatEval. In particular, HateBERT would rank #4 on OffensEval and #6 on HatEval, obtaining the second best F1 score on the positive class.

The portability experiments were run using the best model for each of the in-dataset experiments. Our results show that HateBERT ensures better portability than a generic BERT model, especially when going from generic abusive language phenomena (i.e., offensive language) towards more specific ones (i.e., abusive language or hate speech). This behaviour is expected and provides empirical evidence to the differences across the annotated phenomena. We also claim that HateBERT consistently obtains better representations of the targeted phenomena. This is evident when looking at the differences in False Positives and False Negatives for the positive class, measured by means of Precision and Recall, respectively. As illustrated in Table 4, HateBERT always obtains a higher Precision score than BERT when fine-tuned on a generic abusive phenomenon and applied to more specific ones, at a very low cost for Recall. The unexpected higher Precision of HateBERT fine-tuned on AbusEval and tested on OffensEval 2019 (i.e., from specific to generic) is due to the datasets sharing same data distribution. Indeed, the results of the same model against HatEval support our analysis.

Train	Model	OffensEval 2019		AbusEval		HatEval	
		P	R	P	R	P	R
OffensEval 2019	BERT	–	–	.510	.685	.479	.771
	HateBERT	–	–	.553	.696	.480	.767
AbusEval	BERT	.776	.420	–	–	.545	.571
	HateBERT	.836	.404	–	–	.565	.567
HatEval	BERT	.540	.220	.438	.241	–	–
	HateBERT	.473	.183	.365	.191	–	–

Table 4: BERT *vs.* HateBERT: Portability - Precision and Recall for the positive class. Rows show the dataset used to train the model and columns the dataset used for testing. Best scores are underlined.

4 Conclusion and Future Directions

This contribution introduces HateBERT `base uncased`,[5] a pre-trained language model for abusive language phenomena in English. We confirm that further pre-training is an effective and cheap strategy to port pre-trained language models to other language varieties. The in-dataset evaluation shows that HateBERT consistently outperforms a generic BERT across different abusive language phenomena, such as offensive language (OffensEval 2019), abusive language (AbusEval), and hate speech (HatEval). The cross-dataset experiments show that HateBERT obtains robust representations of each abusive language phenomenon against which it has been fine-tuned. In particular, the cross-dataset experiments have provided (i.) further empirical evidence on the relationship among three abusive language phenomena along the dimension of specificity; (ii.) empirical support to the validity of the annotated data; (iii.) a principled explanation for the different performances of HateBERT and BERT.

[5] HateBERT, the fine-tuned model, and the RAL-E dataset are available at https://osf.io/tbd58/?view_only=d90e681c672a494bb555de99fc7ae780

A known issue concerning HateBERT is its bias toward the subreddit `r/fatpeoplehate`. To address this and other balancing issues, we retrieved an additional ~1.3M messages. This has allowed us to add 712,583 new messages to 12 subreddits listed in Table A.1, and identify three additional ones (`r/uncensorednews`, `r/europeannationalism`, and `r/farright`), for a total of 597,609 messages. This new data is currently used to extend HateBERT.

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.) investigating the connections across the various abusive langauge phenomena.

Acknowledgements

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01—S20049. The author is responsible for the content of this publication.

Ethical Statement

In this paper, the authors introduce HateBERT, a pre-trained language model for the study of abusive language phenomena in social media in English. HateBERT is unique because (i.) it is based on further pre-training of an existing pre-trained language model (i.e., BERT `base-uncased`) rather than training it from scratch, thus reducing the environmental impact of its creation; [6] (ii.) it uses a large collection of messages from communities that have been deemed to violate the content policy of a social media platform, namely Reddit, because of expressing harassment, bullying, incitement of violence, hate, offense, and abuse. The judgment on policy violation has been made by the community administrators and moderators. We consider this dataset for further pre-training more ecologically representative of the expressions of different abusive language phenomena in English than the use of manually annotated datasets.

The collection of banned subreddits has been retrieved from a publicly available collection of Reddit, obtained through the Reddit API and in compliance with Reddit's terms of use. From this collection, we generated the RAL-E dataset. RAL-E will be publicly released (it is accessible also at review phase in the Supplementary Materials). While its availability may have an important impact in boosting research on abusive language phenomena, especially by making natural interactions in online communities available, we are also aware of the risks of privacy violations for owners of the messages. This is one of the reasons why at this stage, we only make available in RAL-E the content of the message without metadata such as the screen name of the author and the community where the message was posted. Usernames and subreddit names have not been used to retrain the models. This reduces the risks of privacy leakage from the retrained models. Since the training material comes from banned community it is impossible and impracticable to obtain meaningful consent from the users (or redditers). In compliance with the Association for Internet Researchers Ethical Guidelines[7], we consider that: not making available the username and the specific community are the only reliable ways to protect users' privacy. We have also manually checked (for a small portion of the messages) whether it is possible to retrieve these messages by actively searching copy-paste the text of the message in Reddit. In none of the cases were we able to obtain a positive result.

There are numerous benefits from using such models to monitor the spread of abusive language phenomena in social media. Among them, we mention the following: (i.) reducing exposure to harmful content in social media; (ii.) contributing to the creation of healthier online interactions; and (iii.) promoting positive contagious behaviors and interactions (Matias, 2019). Unfortunately, work in this area is not free from potentially negative impacts. The most direct is a risk of promoting misrepresentation. HateBERT is an intrinsically biased pre-trained language model. The fine-tuned models that can be obtained are not overgenerating the positive classes, but they suffer from the biases in the manually annotated data, especially for the offensive language detection task (Sap et al., 2019;

[6] The Nvidia V100 GPU we used is shared and it has a maximum number of continuous reserved time of 72 hours. In total, it took 18 days to complete the 2 million retraining steps.

[7] https://aoir.org/reports/ethics3.pdf

Davidson et al., 2019). Furthermore, we think that such tools must always be used under the supervision of humans. Current datasets are completely lacking the actual context of occurrence of a messsage and the associated meaning nuances that may accompany it, labelling the positive classes only on the basis of superficial linguistic cues. The deployment of models based on HateBERT "in the wild" without human supervision requires additional research and suitable datasets for training.

We see benefits in the use of HateBERT in research on abusive language phenomena as well as in the availability of RAL-E. Researchers are encouraged to be aware of the intrinsic biased nature of HateBERT and of its impacts in real-world scenarios.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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. 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.

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. European Language Resources Association.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

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.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 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. Association for Computational Linguistics.

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. 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. 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.

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. Association for Computational Linguistics.

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

J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20):9785–9789.

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. 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. 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.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 446–457. ACM.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and Benchmark Corpora for Hate Speech Detection: a Systematic Review. *Language Resources and Evaluation*, 54(3):1–47.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. Hate speech detection through alberto italian language understanding model. In *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*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

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. 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.

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. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 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.

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. 2019. 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. Association for Computational Linguistics.

Appendix A

Subreddit	Number of posts
apewrangling	5
beatingfaggots	3
blackpeoplehate	16
chicongo	15
chimpmusic	35
didntdonuffins	22
fatpeoplehate	1465531
funnyniggers	29
gibsmedat	24
hitler	297
holocaust	4946
kike	1
klukluxklan	1
milliondollarextreme	9543
misogyny	390
muhdick	15
nazi	1103
niggas	86
niggerhistorymonth	28
niggerrebooted	5
niggerspics	449
niggersstories	75
niggervideos	311
niglets	27
pol	80
polacks	151
sjwhate	10080
teenapers	23
whitesarecriminals	15

A.1: Distribution of messages per banned community composing the RAL-E dataset.

Profanity	Frequency
fucking	52,346
shit	49,012
fuck	44,627
disgusting	15,858
ass	15,789
ham	13,298
bitch	10,661
stupid	9,271
damn	7,873
lazy	7427

A.2: Top 10 profanities in RAL-E dataset.

Pre-processing before pre-training

- 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 replac§ed with a single space;
- extra blank new lines have been removed.

Appendix B

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.

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

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