GRoBERTa: Pre-Trained Embeddings for Offensive Language Detection and Mitigation

No Author Given

No Institute Given

Abstract. We present GRoBERTa (fine-tuned **G**ab **RoBERTa**) - contextual embeddings that are fine-tuned specifically on offensive language, including hate speech and abusive and toxic language. We fine-tune the pre-trained RoBERTa models using a large-scale corpus of posts from a known far-right platform - Gab. Following the success of fine-tuned language models on several natural language tasks, it is our contention that the task of detection and characterization of offensive language would benefit from models that are specifically fine-tuned to the language. To demonstrate, we conduct an extrinsic evaluation of our GRoBERTa embeddings on two downstream tasks of offensive language detection, bias measurement and mitigation. We find that the embeddings are effective in encoding offensive language and can be used to detect, mitigate and analyze such phenomena. We make the embeddings and all our code available to the community to motivate further research in these areas.

Keywords: hate speech · bias mitigation · offensive language.

1 Introduction and Related Work

The proliferation of offensive language, including hate speech and abusive and toxic language, on social media has garnered recent interest from the research community, and understandably so [3,18,21,22,16]. Hate speech, which can be defined as *"language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group,"* [5] can have serious, far-reaching detrimental social effects on the members of the targeted group.

Gab is a censorship-free, social networking platform, prior research has found that the rate of hate speech on Gab is significantly higher than other social networking platforms, (e.g. Twitter) [26]. Analyzing the language used by individuals on such platforms is thus central to understanding, detecting and possibly countering anti-social behaviors including hate speech, abuse and bullying [20]. While most research has focused on how to detect, identify, classify and moderate the use of abusive language online [17,15,8,9], the extent to which language used on a platform such as Gab, proliferates bias and abuse is less well understood. We address this gap by creating specialized contextual representations, so that these can be used in tasks related to hate speech detection and mitigation.

We analyze language use on Gab at the level of word embeddings. We make these embeddings, trained on millions of Gab posts, available to the research community interested in the detection, mitigation and analysis of offensive language.[1] We demonstrate the effectiveness of these embeddings through the following extrinsic evaluations:

1. ***Offensive Language Detection:*** We demonstrate that using GRoBERTA fine-tuned embeddings outperforms the original RoBERTa pre-trained embeddings [13] on detection of offensive language, specifically all three tasks of the OffensEval 2020 challenge [25] and Personal Attack detection task [24].
2. ***Bias Measurement and Mitigation:*** We investigate the measurement and mitigation of biased language in GRoBERTa. To the best of our knowledge, we are the first to measure gender bias in Gab. We also investigate how methods to mitigate extreme gender bias perform on these embeddings.

We use several related terms in this paper, including: hate speech and abusive, offensive and biased language; our contention is that our work is relevant to such anti-social behaviors that manifest on social networking platforms.

Table 1. Descriptive Statistics for Gab data used to create GRoBERTa embeddings

Total Tokens	180 million	Avg. length of a Gab (in words)	18
Total unique tokens	4 million	Avg length of a Gab (in characters)	106
Max. length of a Gab (in words)	951	Min. length of a Gab (in words)	1

2 Data and Preprocessing

The data used in this paper is publicly available online.[2] The corpus consists of all posts made on the Gab platform from June 2017 to May 2018. We filtered this available data for Gabs (i.e. posts on Gab) in the English language, a total of around 9.64 million Gabs (around 50% of the data). We used the language field in the original .json object to determine the language of the Gab. The descriptive statistics of the corpus are shown in Table 1.

To create our embeddings, we chose the RoBERTa model since it been shown to achieve better performance than BERT through training procedure adaptations. We fine-tuned the RoBERTa base model using the masked language model task, as described in the paper by Wolf et al.[23].

3 Experiments and Results

We organize this section as a series of experiments conducted towards extrinsic evaluation of the embeddings and discuss our method, analyses and findings for

[1] link anonymized during review

[2] https://files.pushshift.io/gab/

Table 2. Performance of classification model using GRoBERTa and original RoBERTa embeddings on tasks related to hate speech/offensive language detection.

Task	Accuracy (%)		F1 (%)	
	RoBERTa	GRoBERTa	RoBERTa	GRoBERTa
Offensive Language Detection (Task A)	92.40	**92.72**	91.10	**91.42**
Categorization of Offense Types (Task B)	64.84	**67.01**	52.19	**56.74**
Offense Target Identification (Task C)	80.02	**80.87**	58.32	**59.22**
Personal attack detection	94.35	**94.42**	85.83	**86.23**

each experiment.

Does GRoBERTa improve performance on offensive language detection tasks? Motivated by prior research [11,10], we hypothesize that fine-tuning the language representation models on Gab data would yield better results on tasks like offensive language detection and personal attack detection.

Approach. We use the GRoBERTa embeddings to address classifications tasks from the **OffensEval** 2020 [25] and the task of **detecting personal attacks** in Wikipedia talk pages, proposed by Wulczyn, Thain, and Dixon [24].

The OffensEval contains three tasks: Offensive language Detection (Task A: *Is the text offensive or not offensive?*), automatic Categorization of Offense Types (Task B: *Is the offensive text targeted towards a group or individual or untargeted?*), and Offense Target Identification (Task C: *Who or what is the target of the offensive content?*).

The Personal Attack dataset [24], is another dataset that we used to evaluate our model. This dataset contains labeled discussion comments from English Wikipedia with labels of Not Attack (NOT) and ATTACK.

Results. As shown in Table 2, we find that using GRoBERTa embeddings in the classification model outperforms the original RoBERTa base model on all four tasks on both metrics of accuracy and F1 score.

Do GRoBERTa embeddings encode gender bias? On a platform such as Gab, where abusive language is quite prevalent, gender bias can manifest in the extreme [19].Our hypothesis is that the GRoBERTa embeddings will have a higher degree of bias than more general text.

Approach. To investigate gender bias in GRoBERTa, we compute the direct bias score [2]. Using the results from this computation, we can compare the bias level in GRoBERTa against the level of bias in a set of commonly-used NLP corpora. We use the values calculated for other corpora and benchmark datasets by Babaeianjelodar et al. [1] and show the results in Figure 1.

Results. We compare the Direct Bias level of GRoBERTa against 12 other NLP corpora/benchmarks. As shown in Figure 1, we can see the bias level in GRoBERTa is 0.092, which is the third highest among all the corpora we compare against (a lower score indicates a lower bias level). The bias level in RoBERTa is 0.069, which is also higher than the pre-trained BERT model, which has the

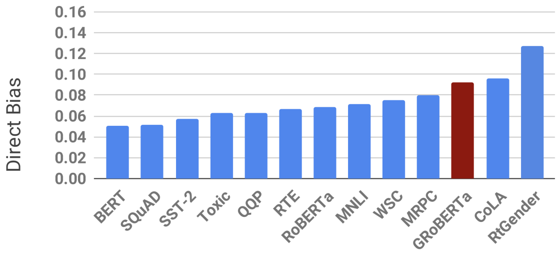

Fig. 1. Comparing the Direct Bias level in commonly-used NLP corpora and GRoBERTa embeddings

Table 3. Effect Size change before and after sentence debiasing using GRoBERTa embeddings. Each row measures binary effect sizes for sentence-level tests, adapted from Caliskan tests. A score closer to 0 represents lower bias.

Test	BERT	GRoBERTa
M/F Names, Career/Family	0.477→ **0.096**	1.390→ **0.074**
M/F Terms, Career/Family	**0.108**→0.194	**0.083**→0.119
M/F Terms, Math/Art	0.253→ **0.194**	**0.420**→0.438
M/F Names, Math/Art	0.399→ **0.075**	1.212→**0.538**
M/F Terms, Science/Art	0.636→ **0.540**	-0.242→**0.055**
M/F Names, Science/Art	1.390→**0.074**	0.943→**0.882**

lowest score. We thus find that fine-tuning on Gab data increases the gender bias in the embeddings greatly by 33.77% (when compared to RoBERTa), due to the nature of the underlying language, which is consistent with our hypothesis. Our results indicate that GRoBERTa embeddings do encode gender bias and do so to a high degree.

How effective are current methods for mitigating bias, when applied to GRoBERTa? Gonen and Goldberg [6] found that the method proposed by Bolukbasi et al. [2] does not adequately capture the gender bias in embeddings and found that even after applying mitigation techniques, gender biases still remain in the embeddings. Given that research on bias mitigation in embeddings is an ongoing endeavour, we propose the use of GRoBERTa to develop mitigation techniques. **Approach.** We investigate how well the state-of-the-art debiasing methods aimed at mitigating bias in contextualized representation models work on GRoBERTa. We choose the debiasing method proposed by Liang et al. [12], which works on the sentence level as opposed to the measure proposed by Bolukbasi et al. [2]. We use pre-defined sentence templates defined in May et al. [14]

Results. To test how well this debiasing method works, we calculate the effect sizes of the test before and after the debiasing [4]. We use the same words, as used in the Caliskan Tests [4], which measure biases in common stereotypes surrounding gendered names with respect to careers, math, and science [7].As shown in Table 3, we find the debiasing technique is able to reduce the bias in 4 out 6 tests proposed by Caliskan et al. [4]. This result is consistent with the results obtained by Liang et al. [12] who experimented with BERT and ELMo embeddings (among others). Their findings for the BERT model are shown in Table 3 for the purposes of comparison. Our findings indicate that the sentence debiasing method works for the majority of the tests, and the GRoBERTa embeddings encode bias in a similar manner as do BERT embeddings.

4 Conclusion

We demonstrate in this paper that a fine-tuned model GRoBERTa can be used as a component for downstream tasks related to offensive language detection and bias mitigation. Our study also shows that Gab exhibits a high degree of gender bias compared to other corpora, ranking the third among 13 commonly-used NLP datasets. As part of future work, we expect to use GRoBERTa to gain perspective into the topics expressed by Gab users when posting about QAnon topics, and report on how the alt-right, echo-chamber like properties of Gab are reflected in the explored topics.

References

1. Babaeianjelodar, M., Lorenz, S., Gordon, J., Matthews, J., Freitag, E.: Quantifying gender bias in different corpora. In: Companion Proceedings of the Web Conference 2020. pp. 752–759 (2020)
2. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in neural information processing systems. pp. 4349–4357 (2016)
3. Brassard-Gourdeau, E., Khoury, R.: Subversive toxicity detection using sentiment information. In: Proceedings of the Third Workshop on Abusive Language Online. pp. 1–10 (2019)
4. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
5. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009 (2017)
6. Gonen, H., Goldberg, Y.: Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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). pp. 609–614 (2019)
7. Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology **74**(6), 1464 (1998)
8. Karan, M., Šnajder, J.: Cross-domain detection of abusive language online. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 132–137 (2018)
9. Kshirsagar, R., Cukuvac, T., McKeown, K., McGregor, S.: Predictive embeddings for hate speech detection on twitter. arXiv preprint arXiv:1809.10644 (2018)

10. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
11. Li, Q., Shah, S., Liu, X., Nourbakhsh, A.: Data sets: Word embeddings learned from tweets and general data. In: ICWSM (2017)
12. Liang, P.P., Li, I.M., Zheng, E., Lim, Y.C., Salakhutdinov, R., Morency, L.P.: Towards debiasing sentence representations. arXiv preprint arXiv:2007.08100 (2020)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
14. May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561 (2019)
15. Mishra, P., Yannakoudakis, H., Shutova, E.: Neural character-based composition models for abuse detection. arXiv preprint arXiv:1809.00378 (2018)
16. Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on bert model. PLOS ONE **15**(8), 1–26 (08 2020). https://doi.org/10.1371/journal.pone.0237861, `https://doi.org/10.1371/journal.pone.0237861`
17. Mubarak, H., Darwish, K., Magdy, W.: Abusive language detection on arabic social media. In: Proceedings of the First Workshop on Abusive Language Online. pp. 52–56 (2017)
18. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. pp. 145–153 (2016)
19. Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y.: A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251 (2019)
20. Waldron, J.: The harm in hate speech. Harvard University Press (2012)
21. Waseem, Z., Davidson, T., Warmsley, D., Weber, I.: Understanding abuse: A typology of abusive language detection subtasks. arXiv preprint arXiv:1705.09899 (2017)
22. Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: 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). pp. 602–608 (2019)
23. Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45 (2020)
24. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1391–1399 (2017)
25. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, c.: SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In: Proceedings of SemEval (2020)
26. Zannettou, S., Bradlyn, B., Cristofaro, E.D., Sirivianos, M., Stringhini, G., Kwak, H., Blackburn, J.: What is gab? A bastion of free speech or an alt-right echo chamber? CoRR **abs/1802.05287** (2018), `http://arxiv.org/abs/1802.05287`