Counter hate speech detection in Youtube conversations*

Pedro Fialho[1], Ricardo Ribeiro[2,3], Fernando Batista[2,3], Gil Ramos[1], António Fonseca[1], Sérgio Moro[1,4], Rita Guerra[5], Paula Carvalho[3], Catarina Marques[6], and Cláudia Silva[7]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[3] INESC-ID, Lisboa, Portugal
[4] University of Jordan, Amman, Jordan
[5] ISCTE-Instituto Universitário de Lisboa and Center for Psychological Research and Social Intervention (CIS-ISCTE), Lisboa, Portugal
[6] ISCTE-Instituto Universitário de Lisboa and Business Research Unit (BRU-ISCTE), Lisboa, Portugal
[7] ITI-LARSyS and IST, Lisboa, Portugal

Abstract. One of the forms to fight the spread of hate speech is to reply to such utterances with counter speech. In this paper, we present models to identify counter speech, based on an annotated set of comments to Youtube videos spoken in Portuguese. We leverage the sequence of replies to comments, to form a corpus of pairs of comments, where a target is labelled as neutral, hate speech or counter speech, relative to a context, which corresponds to a preceding comment. To the best of our knowledge, this is the first corpus with counter speech examples in Portuguese. Using such corpus, we compute models by fine-tuning pre-trained models based on Transformers, and experiment with both multilingual and Portuguese pre-trained models. Our approach follows a recent work for English, both in corpus design and experimental setup, and we obtain similar performance results in Portuguese.
Warning: This work contains offensive and hateful text that some might find upsetting. It does not represent the views of the authors.

Keywords: counter hate speech · hate speech · social networks · transformers · text classification.

1 Introduction

Hate speech has prevailed in online discussion platforms [20]. Replies to hate speech comments that provide an informed interpretation of the hateful content, towards countering it, are known as counter speech [4], and are one of the forms to prevent the spread of hate speech [12]. Counter speech detection is a

* This work was funded in part by the European Union under Grant CERV-2021-EQUAL (101049306).

recent research topic. As counter speech is rare, and online discussion platforms may limit streamlined access to public discussions from their users, automatic detection enables, for instance, to enhance the scale and speed of data collection, also across languages [19].

Our aim is to detect counter speech in Portuguese, using pre-trained BERT-based models [8], which rely on Transformers [25] and have been successfully employed, performance wise, in diverse Natural Language Processing tasks. We follow [27], where counter speech in English was best detected from a pair of utterances, and propose a method to form a corpus from comments found in the discussion section of Youtube[8] videos. Each example represents whether a target comment is neutral, hate speech or counter speech, relative to a context comment that precedes it in the discussion. The machine learning framework that supports our experiments is also equivalent to that of [27], but we use pre-trained models for Portuguese instead.

We compute a model with the English corpus from [27], and obtained results similar to those originally reported in such work. We then compute a model with our Portuguese corpus, using the same framework, and also obtained similar results to those of our English model. As such, our Portuguese model achieves competitive performance, although results between languages are not comparable, since the corresponding corpora is different and have significant differences in size. This paper also presents the performance obtained with various pre-trained models suitable for Portuguese, such as the multilingual BERT [8] or a specialized model of hate speech in Portuguese [18].

Our main contribution is a Portuguese corpus suitable to train and evaluate models on counter speech detection. To the best of our knowledge, this is the first corpus focusing on counter speech examples in Portuguese. As the labels in our corpus were manually annotated, and counter speech annotation is prone to misjudgments [3], we also provide an in-dept discussion of the predictions of our models for a selection of test examples, highlighting the intricacies of counter speech detection.

This paper is organized as follows: Section 2 explores existing literature and methods for counter speech detection, setting a foundation for our research. Section 3 introduces the corpus, describing its creation and the rationale behind its structure. The methodology for adapting and fine-tuning Transformer-based models to our corpus is elaborated in Section 4, followed by Section 5 which presents the outcomes of our experiments. Section 6 shows an analysis of the errors encountered, providing insights into the limitations and challenges of our models. Section 7 concludes with a summary of our contributions and a look towards future research directions, underscoring the potential of our work to advance the detection of counter speech within Portuguese-speaking online communities.

[8] https://www.youtube.com/

2 Related work

Most works focused on counter speech span into studies for generative approaches, typically with the aim of automatically replying to hate speech [28], and detection approaches. We focus on the latter, which has been employed, for instance, in support of data collection for languages which are under-resourced in counter speech annotations [19].

Current approaches for counter speech detection typically rely on Transformer-based models, such as BERT [8], which are pre-trained on generic data, and fine-tuned on corpora featuring counter speech examples [27]. However, early works obtained similar performance from models based on traditional machine learning or early neural networks [17,21,11].

Suitable corpora for counter speech detection models has been sourced from user inputs on various social media platforms, such as Twitter [19] and Reddit [21,27]. Also, machine generated corpora containing counter speech examples has been produced, for instance leveraging modern generative models by using prompt engineering [2].

We rely on the Youtube platform to build a counter speech detection corpus from user comments in Portuguese, which leverages the conversations occurring as replies to comments. Previous research has explored counter speech detection on YouTube for English comments, but it was limited to direct responses to videos to reduce the likelihood of off-topic discussions [17].

For a more curated collection of counter speech, the CONAN corpus [5] contains examples formulated by experts, in English, French and Italian, thus featuring linguistically and semantically verified instances of counter speech. This corpus employs paraphrasing to augment the number of pairs per language, and includes annotations for additional information such as counter speech type. Additional corpora has spanned from the CONAN corpus design, to focus on related aspects of hate and counter hate speech, such as hate targets [9] and background knowledge [6].

Examples in corpora designed for counter speech detection are typically formed by a label and either a single utterance [17], or a pair of utterances, where one corresponds to the label and the other is the context that supports such labelling [27]. Labels typically describe a binary classification task, where the utterance is labelled as counter speech or not [17], or a multi-label classification task, where the corpora also contains examples for other hate related phenomena, such as hate speech [27] or different types of counter speech [17,13].

Given that labelling a text as counter speech is prone to subjective judgments influenced by the background of the annotator [3], and naturally occurring counter speech is rare, some works opt for bulk or machine-based annotation, trading a loss in annotation accuracy for an increase in corpora size [11,19]. Some of these approaches rely on lists of hate related words to filter conversations [21,27,19]. However, some works invest in the production of more relevant and accurate examples, for instance by enforcing grammatical correctness [29], enabling parametric generation of counter speech examples [22] or annotation by experts [5].

Multilingual approaches to counter speech detection are still scarce, but corpora exists for various languages. Our work contributes a new corpus on counter speech detection, based on Youtube comments written in Portuguese. To the best of our knowledge, this is the first corpus with counter speech examples in Portuguese, although corpora with hate speech examples in Portuguese exists from previous works [10,15,24].

3 Paired comments corpus

We built a corpus from a collection of comments, to Youtube videos spoken in Portuguese, manually labelled as hate speech, counter speech or none, under the scope of the project *kNOwHATE: kNOwing online HATE speech*[9], which rules the availability of this data. The design of our corpus is the same as that of the corpus introduced by [27], to address the same task in English, based on these labels.

A group of four annotators, with backgrounds in language sciences and social psychology, was involved in labelling our collection. For a subset of this collection, to each comment correspond four labels, one for each of the four annotators. For the remainder of the collection, to each comment corresponds a single annotation, from one of the four annotators. Following [27], we consider the latter as the silver set, intended only to train our models, and the former as the gold set, intended to evaluate the performance of our models, since the definite label for a comment is based on four labels, and is hence more reliable. We select the label where most annotators agree as the definite label for a comment.

In YouTube, some comments are follow-up replies to other comments, describing a conversation triggered by a comment to the video. We leverage this structure to form an example with a target comment, which is being labeled, and the context for that target, which is one of the preceding comments. We only consider comments with follow-up replies, to ensure that a context comment is always available. Fig. 1 presents an example from our corpus.

previous context: *Acabei de ver nesse programa, a propaganda do veganismo e a agenda de forçar os homens a consumir soja. [I just saw on that show, the vegan propaganda and the agenda of forcing man to ingest soybeans]*
target: *Não há nenhum estudo que comprove que o consumo de soja está ligado ao efeminismo no homem. Por isso não fale mentiras propagadas pela midia. [there is no study proving that soybeans consumption is linked to an effemination of man. So don't talk lies spread by media]*
label: *counter speech*

Fig. 1. Example of counter speech, extracted from our corpus.

[9] https://knowhate.eu/

We consider that the context in a counter speech example of our corpus is always a comment labelled as hate speech in the source collection. Moreover, a counter speech comment preceded by multiple hate speech comments yields multiple counter speech examples, with the context set to each of the hate speech comments, and the target set to the same counter speech comment. For the neutral and hate speech examples of our corpus, the context is always the first comment, which started the follow-up sequence of replies. Using this method, we computed 298 examples from 825 comments in the gold set, and 8272 examples from 23912 comments in the silver set. In the following section we define our setup to assess the performance of a model based on this examples, where the gold examples are employed as test set, while the silver examples are further split in train and validation partitions, with the latter corresponding to 10% of the silver examples. Label and example distribution is shown in Table 1.

Table 1. Number of examples in our corpus, per label and partition.

Partition	**Neutral**	**Counter**	**Hate**	**Total**
train	2080	739	4625	7444
validation	232	82	514	828
test	101	28	169	298

4 Experimental setup

Our experiments are based on a neural network where pre-trained models based on BERT [8] are fine-tuned to our corpus. We follow the experimental setup of [27], but instead using pre-trained models suitable for Portuguese.

As such, our neural network is composed by a pre-trained BERT-based model, of which we select the weights for the classification/CLS token [8] to forward to a fully connected layer, with 768 neurons and Tanh activation. We then apply dropout, and follow with a final fully connected layer, with softmax activation and 3 neurons, corresponding to the 3 labels we aim to predict. As optimizer we employ Adam [14], with learning rate set to 0.00001. Both dropout and optimizer operate at the same rates as in all experiments by [27]. All our models are trained for 10 epochs with a batch size of 16, following the accuracy results on train and validation sets, shown in Fig. 2.

We report results for various instances of this setup, which only differ in the underlying pre-trained model, such that each employs one of to the following BERT-based models:

– the base version of the XLM-RoBERTa model [7], henceforth mentioned as *xlm-roberta*

– the base and cased version of the multilingual BERT model [8], henceforth named *bert-b-c-multi*
– a specialized model of hate speech in Portuguese [18], henceforth named *hate-bert-tuga*
– the Portuguese BERT model BERTimbau[23], pre-trained in Brazilian corpora, henceforth named *bertimbau*

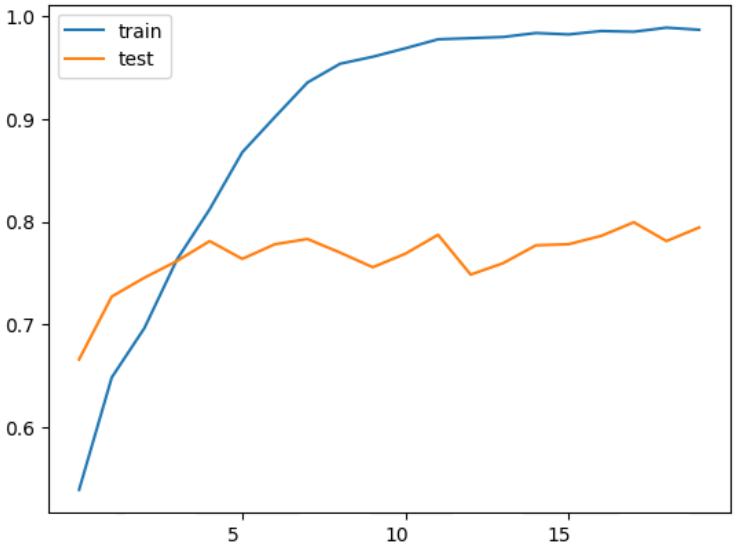

Fig. 2. Accuracy results on train and validation sets, along training our neural network for 20 epochs. The horizontal axis corresponds to number of epochs, while the vertical axis corresponds to accuracy scores.

All our experiments are run on Google Colab[10], using the Transformers library [26] and Tensorflow [1] for model manipulation.

5 Results

To validate our implementation of the experimental setup, we first build a model with the English corpus and pre-trained model employed by [27], henceforth named *reEN*, and assess its performance. The only difference from our setup for Portuguese is the pre-trained model, which for English is the RoBERTa model [16].

In Table 2 we present the performance results for our Portuguese models, and for two English models, namely one of the models reported in [27], here named as the *EN* model, and our replication of it, here named the *reEN* model.

[10] https://colab.research.google.com/

The results for the *EN* model are those reported in [27] for the simplest model trained on a pair of utterances, which we followed in designing our Portuguese models. The *reEN* model is our implementation of the *EN* model, using the same setup of our Portuguese models, but with English data from [27], as obtained from https://github.com/xinchenyu/counter_context

Table 2. F1 scores for the neutral, counter speech and hate speech classes, and their weighted average, with best results shown in boldface. The first two rows correspond to models for English, while the remaining correspond to models for Portuguese.

Model	**Neutral**	**Counter**	**Hate**	**Average**
EN	0.70	0.44	0.59	0.61
reEN	0.694	0.470	0.559	0.607
bert-b-c-multi	0.670	0.410	0.753	0.693
xlm-roberta	**0.762**	0.386	0.772	**0.732**
hate-bert-tuga	0.652	0.326	**0.786**	0.697
bertimbau	0.720	**0.417**	0.764	0.717

For English, our *reEN* model achieves similar performance to that of the *EN* model, which it replicates. As such, we assure that our implementation, which we employ in both English and Portuguese models, corresponds to that of [27].

For Portuguese, the multilingual *xlm-roberta* model achieved the best overall performance, which we consider as the average for the performances of all classes, weighted by the number of instances on each class, as reported in Table 2. The second best model, considering average performance, is the Portuguese-specific *bertimbau* model, although it achieves the best performance of all models on detecting instances of counter speech. The *bert-b-c-multi* model achieved the worst overall performance, but the second best performance on counter speech detection. The *hate-bert-tuga* model achieved the second worst overall performance, but the best performance on detecting instances of hate speech, which was expected since it is tailored for hate speech data.

6 Error analysis

Of the 298 test examples, all of our models correctly predicted the label on 132 examples, and failed to predicted the correct label for 25 test examples. One of the examples that all models failed to correctly predict the label is shown in Fig. 3. All models predicted this example as hate speech.

In Table 3 we present an overview for the amount of examples that only each model was able to correctly classify, both per class and in total, and in the following we present a selection of such examples.

context: *Tudo apoiantes do lula livre que o bloco de esterco anda a importar para cá. [All supporters of "Lula Livre" that the block of manure is importing here.]*
target: *Sim, têm um presidente ditador. E se houver muita gente a pensar como o senhor, aqui acontecerá a mesma coisa. [Yes, they have a dictator president. And if there are many people thinking like you, the same thing will happen here.]*
label: counter speech

Fig. 3. Test example where all models failed to correctly predict the label.

Table 3. Number of examples that only a certain model was able to correctly classify, per class and in total.

Model	Neutral	Counter	Hate	Total
bert-b-c-multi	3	1	3	7
xlm-roberta	4	5	3	12
hate-bert-tuga	1	0	7	8
bertimbau	1	0	2	3

The *hate-bert-tuga* model was the only to correctly predict the label on 8 test examples, and none of these are counter speech examples. One of such examples is shown in Fig. 4, which all other models predicted as neutral.

context: *As brasileiras são muito trabalhadoras. Pena não fazerem descontos . Lixo . [Brazilian women are very hardworking. It's a shame they don't offer discounts. Trash]*
target: *hahahhahahahahahahahhahahahahahhahahaha o melhor comentário [hahahhahahahahahahahhahahahahahhahahaha the best comment]*
label: hate speech

Fig. 4. Example where only *hate-bert-tuga* was able to correctly predict the label.

The *bert-b-c-multi* model was the only to correctly predict the label on 7 examples, and only one is counter speech. Such example is shown in Fig. 5, which all other models predicted as hate speech.

The *bertimbau* model was the only to correctly predict the label on 3 examples, and none of these are counter speech examples. One of such examples is shown in Fig. 6, which all other models predicted as neutral.

The *xlm-roberta* model was the only to correctly predict the label on 12 examples, one of which is shown in Fig. 7. All other models predicted this example as hate speech.

context: *os nossos escravos sao rebeldes [our slaves are rebels]*
target: *Somos a 9(nona) economia mais forte do planeta é temos a moeda mais forte da américa latina. A colónia superaria o colonizador se portugal não estivesse na zona do euro. [We are the 9th (ninth) strongest economy on the planet and we have the strongest currency in Latin America. The colony would surpass the colonizer if Portugal were not in the euro zone.]*
label: counter speech

Fig. 5. Example where only *bert-b-c-multi* was able to correctly predict the label.

context: *Eu aqui em Moçambique so nao me fazem a folha porque a 38 anda sempre na cintura....forca por ai eu assim que poder ir a portugal vou estar ao teu lado Mário FORÇA [Here in Mozambique, they only don't overpower me because the 38 is always around my waist....hang in there, as soon as I can go to Portugal I'll be by your side Mário STRENGTHEN]*
target: *Quando vieres diz algo amigo. Esse é o meu telegram M1143.. Um grande abraço e muita força aí [When you come around, say something friend. This is my telegram M1143.. A big hug and lots of strength there]*
label: hate speech

Fig. 6. Example where only *Bertimbau* was able to correctly predict the label.

7 Conclusion

This research has made notable progress in detecting hate speech online, focusing particularly on the Portuguese-speaking segment of YouTube. We have created and validated a unique corpus of comments in Portuguese, specifically tailored to identify counter speech effectively.

Our exploration involved fine-tuning advanced Transformer-based models, utilizing both multilingual and Portuguese-specific pre-trained models. This approach, inspired by similar endeavors in English, allowed us to not only replicate but also extend the existing framework to accommodate the nuances of Portuguese. The performance metrics of our models indicate that they are on par with their English counterparts, suggesting that the techniques for detecting counter speech are effective across different languages, at least within the context of languages covered by the pre-trained models employed.

Future work includes designing and experimenting variations to the neural architecture, more in-depth analysis of examples between models, and exploring the usage of machine generated corpora, although the latter is mostly available for English [2].

Our corpus is made only from comments with follow-up replies, as such all single comments are discarded, although these are annotated in our source collection. Future work also includes pairing single comments, for instance using the title of the video as context. Finally, we also aim to enhance the credibility of our findings by incorporating statistical tests, like the t-test, to confirm the significance of our results.

context: *Afinal não é gay?? oh,que pena, eu a pensar que iamos ter a Sara Carbonero livre no nosso team. [After all, he's not gay?? oh, what a shame, I thought we would have Sara Carbonero free on our team.]*
target: *É tão triste falares de orientaçoes sexuais como "teams" Mas pronto era uma piada, de certeza que não sentes isso. [It's so sad that you talk about sexual orientations as "teams" But then it was a joke, I'm sure you don't feel that way.]*
label: counter speech

Fig. 7. Example where only *xlm-roberta* was able to correctly predict the label.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Ashida, M., Komachi, M.: Towards automatic generation of messages countering online hate speech and microaggressions. In: Narang, K., Mostafazadeh Davani, A., Mathias, L., Vidgen, B., Talat, Z. (eds.) Proceedings of the WOAH 2022 Workshop. pp. 11–23. ACL, Seattle, Washington (Hybrid) (Jul 2022). https://doi.org/10.18653/v1/2022.woah-1.2, https://aclanthology.org/2022.woah-1.2
3. Baumler, C., Sotnikova, A., Daumé III, H.: Which examples should be multiply annotated? active learning when annotators may disagree. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the ACL 2023. pp. 10352–10371. ACL, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.658
4. Cepollaro, B., Lepoutre, M., Simpson, R.M.: Counterspeech. Philosophy Compass **18**(1), e12890 (2023). https://doi.org/https://doi.org/10.1111/phc3.12890
5. Chung, Y.L., Kuzmenko, E., Tekiroglu, S.S., Guerini, M.: CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the ACL. pp. 2819–2829. ACL, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1271
6. Chung, Y.L., Tekiroğlu, S.S., Guerini, M.: Towards knowledge-grounded counter narrative generation for hate speech. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the ACL: ACL-IJCNLP 2021. pp. 899–914. ACL, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.79
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the ACL 2020. pp. 8440–8451. ACL, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747

8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the NAACL 2019: HLT, Volume 1. pp. 4171–4186. ACL, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
9. Fanton, M., Bonaldi, H., Tekiroğlu, S.S., Guerini, M.: Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of ACL 2021 and IJCNLP 2021 (Volume 1: Long Papers). pp. 3226–3240. ACL, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.250
10. Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., Nunes, S.: A hierarchically-labeled Portuguese hate speech dataset. In: Roberts, S.T., Tetreault, J., Prabhakaran, V., Waseem, Z. (eds.) Proceedings of the Third Workshop on Abusive Language Online. pp. 94–104. ACL, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3510
11. Garland, J., Ghazi-Zahedi, K., Young, J.G., Hébert-Dufresne, L., Galesic, M.: Countering hate on social media: Large scale classification of hate and counter speech. In: Akiwowo, S., Vidgen, B., Prabhakaran, V., Waseem, Z. (eds.) Proceedings of the WOAH 2020 Workshop. pp. 102–112. ACL, Online (Nov 2020). https://doi.org/10.18653/v1/2020.alw-1.13
12. Garland, J., Ghazi-Zahedi, K., Young, J., Hébert-Dufresne, L., Galesic, M.: Impact and dynamics of hate and counter speech online. EPJ Data Sci. **11**(1), 3 (2022). https://doi.org/10.1140/EPJDS/S13688-021-00314-6, https://doi.org/10.1140/epjds/s13688-021-00314-6
13. Gupta, R., Desai, S., Goel, M., Bandhakavi, A., Chakraborty, T., Akhtar, M.S.: Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the ACL 2023 (Volume 1: Long Papers). pp. 5792–5809. ACL, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.318
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
15. Leite, J.A., Silva, D., Bontcheva, K., Scarton, C.: Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In: Wong, K.F., Knight, K., Wu, H. (eds.) Proceedings of the AACL 2020 and IJCNLP 2020. pp. 914–924. ACL, Suzhou, China (Dec 2020), https://aclanthology.org/2020.aacl-main.91
16. 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. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
17. Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S.K., Goyal, P., Mukherje, A.: Thou shalt not hate: Countering online hate speech. In: Thirteenth International AAAI Conference on Web and Social Media (2019)
18. Matos, B.C.: Automatic Hate Speech Detection in Portuguese Social Media Text. Master's thesis, Instituto Superior Técnico (Nov 2022)
19. Möhle, P., Orlikowski, M., Cimiano, P.: Just collect, don't filter: Noisy labels do not improve counterspeech collection for languages without annotated resources. In: Chung, Y.L., Bonaldi, H., Abercrombie, G., Guerini, M. (eds.) Proceedings of

the 1st Workshop on CounterSpeech for Online Abuse (CS4OA). pp. 44–61. ACL, Prague, Czechia (Sep 2023), https://aclanthology.org/2023.cs4oa-1.4
20. Mondal, M., Silva, L.A., Benevenuto, F.: A measurement study of hate speech in social media. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. p. 85–94. HT '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3078714.3078723
21. Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y.: A benchmark dataset for learning to intervene in online hate speech. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the EMNLP-IJCNLP 2019. pp. 4755–4764. ACL, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1482
22. Saha, P., Singh, K., Kumar, A., Mathew, B., Mukherjee, A.: Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 5157–5163. International Joint Conferences on Artificial Intelligence Organization (7 2022), https://doi.org/10.24963/ijcai.2022/716, aI for Good
23. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear) (2020)
24. Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., Benevenuto, F.: HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) Proceedings of the LREC 2022. pp. 7174–7183. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.777
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
26. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the EMNLP 2020: System Demonstrations. pp. 38–45. ACL, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6
27. Yu, X., Blanco, E., Hong, L.: Hate speech and counter speech detection: Conversational context does matter. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the NAACL 2022: HLT. pp. 5918–5930. ACL, Seattle, United States (Jul 2022), https://aclanthology.org/2022.naacl-main.433
28. Zheng, Y., Ross, B., Magdy, W.: What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In: Chung, Y.L., Bonaldi, H., Abercrombie, G., Guerini, M. (eds.) Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA). pp. 62–71. ACL, Prague, Czechia (Sep 2023), https://aclanthology.org/2023.cs4oa-1.5
29. Zhu, W., Bhat, S.: Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the ACL: ACL-IJCNLP 2021. pp. 134–149. ACL, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.12, https://aclanthology.org/2021.findings-acl.12