Just Collect, Don't Filter: Noisy Labels Do Not Improve Counterspeech Collection for Languages Without Annotated Resources

Pauline Möhle, Matthias Orlikowski and **Philipp Cimiano**
Center for Cognitive Interaction Technology (CITEC)
Bielefeld University

Abstract

Counterspeech on social media is rare. Consequently, it is difficult to collect naturally occurring examples, in particular for languages without annotated datasets. In this work, we study methods to increase the relevance of social media samples for counterspeech annotation when we lack annotated resources. We use the example of sourcing German data for counterspeech annotations from Twitter. We monitor tweets from German politicians and activists to collect replies. To select relevant replies we a) find replies that match German abusive keywords or b) label replies for counterspeech using a multilingual classifier fine-tuned on English data. For both approaches and a baseline setting, we annotate a random sample and use bootstrap sampling to estimate the amount of counterspeech. We find that neither the multilingual model nor the keyword approach achieve significantly higher counts of true counterspeech than the baseline. Thus, keyword lists or multilingual classifiers are likely not worth the added complexity beyond purposive data collection: Already without additional filtering, we gather a meaningful sample with 7,4% true counterspeech.

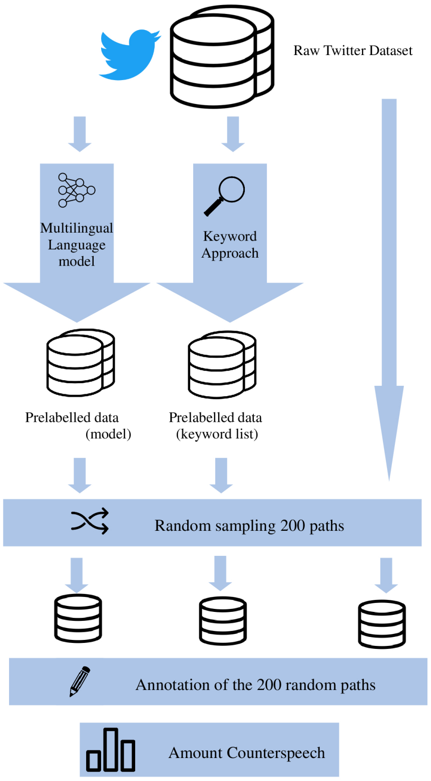
Raw Twitter Dataset
Multilingual Language model
Keyword Approach
Prelabelled data (model)
Prelabelled data (keyword list)
Random sampling 200 paths
Annotation of the 200 random paths
Amount Counterspeech

Figure 1: Overview of the Experimental Setup

1 Introduction

Abusive speech is a serious problem on social media, causing harm, division, and offline violence (Benesch, 2014)[1].

While bans are effective, they have two drawbacks - apart from the fact that they can restrict freedom of speech (Gagliardone et al., 2015) - namely that deleting abusive posts would not tackle the entire problem of online abusiveness because "only the most egregious forms of hate speech [...] are generally considered unlawful" (Izsák, 2015) and that banning abusive speech "can miss out on how societies evolve through [...] disagreement" because it "can [...] be thought of as a window into deeply rooted tensions and inequalities" (Gagliardone et al., 2015). On these grounds, many organisations have started calls for the civil online community to counter and marginalise hateful and abusive messages. Today, online counterspeech is an established practice with country-specific organisations in many national contexts.

In contrast, annotated datasets are available only for a few languages (Mathew et al., 2019; Chung et al., 2021; Goffredo et al., 2022), which restricts computational analyses using Natural Language Processing. When trying to bridge this gap by building datasets for additional languages, we are

[1] Content Notice: The appendix includes an error analysis with examples of abusive language, in particular slurs and anti-muslim racism.

Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), pages 44–61
September 11, 2023 ©2023 Association for Computational Linguistics

confronted with specific challenges. In particular, as counterspeech is rare on social media, it is not clear how to best collect a dataset that is relevant for annotation, i.e. which has a reasonable chance of containing actual examples of counterspeech. While prior work uses different approaches to gather data and sometimes discusses their limitations, counterspeech data collection so far has not been studied explicitly.

In this work, **we evaluate whether filtering based on automatically-assigned labels can help to increase the amount of counter speech in a data sample *when we do not have annotated data available in the language*** we work with (see overview 1). The setting for this evaluation is the creation of a German counterspeech dataset. We study two approaches which assign provisional labels to tweets: a) a keyword-based approach to detect abusive speech as a precondition for counterspeech replies and b) a multilingual language model fine-tuned on an English counterspeech dataset. As a baseline, we also take a random sample from the original dataset without applying any additional filtering. All three samples are manually annotated by the first author for abusive speech, counterspeech, and neutral speech using definitions by Vidgen et al. (2021). To derive estimates of the counter speech counts in the original datasets, we use bootstrap sampling with the annotated samples. For our evaluation we collect an initial dataset of replies to German-speaking political public figures. We use the provisional labels to filter the dataset and draw a random sample for each approach.

Both evaluated filter approaches represent tradeoffs between resource availability and reasonable effort motivated by the goal of increasing data relevance for annotation. That is, we want to increase the amount of true counterspeech in a sample, not necessarily develop the most accurate classifiers. For example, it is acceptable if our multilingual model has mediocre performance on English text as long as it helps to increase the German sample's amount of counter speech as part of the complete setup.

We find that both the multilingual model and the keyword approach achieve higher boostrap mean counts of true counterspeech than the baseline. However, confidence intervals overlap, so it is not likely that these gains generalize beyond our samples. Nevertheless, already without additional filtering, our data collection produces a meaningful sample with 7,4% true counterspeech. Thus, data collection efforts should weigh the added complexity against unlikely benefits. If filter methods are to be used and the additional effort seems justified, methods need to be improved substantially over the presented approaches.

2 Related Work

As investigating the dynamics and types of counterspeech has found increasing interest, there is a variety of approaches to data collection. One line of work has investigated methods to source examples from study participants, either from experts (Chung et al., 2019) or the crowd (Qian et al., 2019), also by means of post-editing the output of text generation models (Tekiroğlu et al., 2020; Fanton et al., 2021). While these works are focused on collecting high-quality examples for counterspeech generation, we want to collect counterspeech from social media to improve classification and analyses of naturally occurring counterspeech.

Work which gathers data from social media falls into two categories. The first category collects examples based on *identifying counterspeech communities*. Examples are Goffredo et al. (2022) and Procter et al. (2019) who retrieve tweets of accounts known to produce counterspeech or encourage other users to do so and collect these replies. Similarly, Vidgen et al. (2021) identify specific communities to find abusive speech and counterspeech in the conversations. These approaches inform our initial data collection (see Section 3). The second category is based on *abusive content as trigger for counterspeech*. Mathew et al. (2019) built a dataset by manually searching for Youtube videos with abusive titles to collect counterspeech in the comments below. In a similar way, Albanyan and Blanco (2022) use abusive tweets from an existing corpus to query Twitter for additional abusive content to then annotate the replies to the abuse in expectation of counterspeech. Yu et al. (2022) use a keyword list to find potential abusive content, while Mathew et al. (2018) and Mathew et al. (2020) use templates of abusive phrases, all assuming to potentially find counterspeech among replies to abusive posts. Counterspeech collection from abusive content inspires one of the filter approaches (based on abusive keywords) which we evaluate (see Section 4). He et al. (2020) use an approach spanning both mentioned categories with a keyword list for both abusive speech and counterspeech.

Although there has been significant focus on abusive speech in German (Bretschneider and Peters, 2017; Bai et al., 2018; Wiegand et al., 2018; Struss et al., 2019; Risch et al., 2021), there has been limited exploration of counterspeech. As a notable exception, Garland et al. (2020) build a large German counterspeech dataset by using a unique period in German social media where users labeled themselves as proponents of abusive speech (i.e., extreme right-wing group) or counterspeech. However, due to this distant supervision the labels are less accurate. Also, the dataset is not publicly available to protect the privacy of users in this particularly sensitive context (see also Garland et al.).

3 Data Collection

To collect data for our study, we chose Twitter as a platform due to its accessibility (at the time of data collection). We first qualitatively explored conversations around political German-speaking Twitter accounts. A key observation was that abusive speech usually does not appear in original tweets but in replies. Here, users often attack the author of the original tweet, other users in the conversation or social groups which are mentioned in the initial tweet. Therefore we selected 12 seed accounts which receive a lot of replies to their tweets, often containing abusive speech. These seed accounts include politicians, journalists and activists of different genders, ages and political views (see App. B.1).

During the time of data collection (November 2022), the Twitter API readily allowed to stream data in real time. The level of API access used for this work allowed us to monitor the 12 most recent posts of all seed accounts. If the data collection was interrupted, e.g., due to network errors, we filled the gaps by retrieving missing tweets via the non-streaming API. If a gap could not be filled, the corresponding path of replies was deleted entirely. Data was being collected continuously from the 15th to the 25th of November 2022.

We divided the collected dataset into the individual paths of replies to each original tweet. That is, if we think of all direct replies to a tweet - a conversation in Twitter's terms - as a tree with the original tweet as the root node, we are interested in all the root-to-leaf paths. To produce these paths, we used the ConvoKit library (Chang et al., 2020).

As we follow a definition of counterspeech as a reply to an abusive statement (see Section 5.1), relevant conversation paths need to have a minimum length of two. However, if the original tweet is not abusive, the second reply is the earliest post in the conversation that can contain counterspeech. Therefore, we manually annotate the root tweets in our dataset as being abusive or neutral. Of the 188 root tweets 21 were labelled as abusive speech and 167 were labeled as neutral. Based on these annotations, we included conversation paths with a minimum length of three for neutral roots and with a minimum length of two for abusive roots.

The final dataset contains 85,942 unique tweets divided into 48,550 conversation paths, where one tweet can be included in multiple conversation paths. All these paths go back to the 188 root tweets posted by the 12 seed accounts during the time of data collection.

4 Filter Approaches

We study two different filter techniques to make the dataset more relevant in terms of counterspeech: one approach based on German abusive keywords and another approach based on a multilingual language model fine-tuned on an English counterspeech dataset.

The **keyword** approach is based on the idea that if we identify an abusive tweet in a conversation path, the path meets at least a necessary condition to also contain counterspeech. We use the keyword list of abusive terms by Bai et al. (2018). Nine words were deleted from the list after preliminary experiments showed that they mostly produce erroneous matches, retaining 11,295 keywords. When searching for matches, we ignore casing and only accept verbatim matches of complete tokens (no inflections, no partial matches). When filtering the dataset, we include paths where at least one tweet contains an abusive keyword, unless the only match is in the last tweet. Full counts of most-frequent matches are in App. A.2.

The second approach, based on a **multilingual counterspeech classifier**, is inspired by Chung et al. (2021) who trained a multilingual model for counterspeech type classification in English, Italian and French. As there are no counterspeech detection annotations available for German but for other languages, training a multilingual model is a promising option to train a classifier. As base model we use XLM-T (Barbieri et al., 2022), a multilingual Twitter language model derived from XLM-RoBERTa (Conneau et al., 2020). For fine-tuning, we use an English Reddit dataset (Yu et al., 2022) with a total of 5,000 comments for training, 1132 comments for validation and 713 comments for testing. Although Reddit data is of course different from Twitter data because different platform affordances and cultures, using the dataset is a reasonable trade-off for a number of reasons: Yu et al. (2022) also use the counterspeech definition by Vidgen et al. (2021) which we base our evaluation on. The dataset represents a situation of language use similar to our dataset as it contains natural social media speech with replies to other users (that is, not isolated posts). Crucially, the dataset is comparatively large and also contains a neutral class, so that abusive speech, counterspeech as well as neutral speech are represented in proportions which seem appropriate for our inital dataset. Also, as the base language model XLM-T was pre-trained on Twitter data, it should still represent our target domain Twitter well after just a few epochs of fine-tuning on Reddit data. In preliminary experiments we select the configuration with the highest F1 (0.45) for counterspeech on the English test set, accepting lower performance for abusive and neutral speech (overall F1 of 0.59, see further details on data, training and evaluation in App. A.1). To filter the dataset, we first use the trained model to classify all tweets in the initial dataset. Then, we select all conversation paths with at least one comment classified as counterspeech.

5 Evaluation

To evaluate the two filtering methods, we manually annotate samples for each method and compare the amount of counter speech produced by each method[2]. We take random samples of 200 paths from the filtered datasets. Additionally, we create a random sample of 200 paths from the original unfiltered dataset as a baseline. We sample complete conversation paths to ensure in-context annotation (see Section 5.1). In turn, as paths have different lengths, we accept that the evaluation samples contain different numbers of tweets. The baseline sample has 810 tweets, the keyword sample 1,816 tweets and the sample of the multilingual classifier 927 tweets.

For a more meaningful comparison between the samples, we do not compare directly but resample to derive estimates of counter speech counts. We use bootstrap sampling (Efron, 1979) with a sample size of 810, that is, the size of the smallest evaluation sample. We draw 1,500 bootstrap samples and compute the mean and 95% confidence intervals for the number of examples labelled as counterspeech.

5.1 Human Annotation

The data were manually annotated by the first author using the annotation tool *LightTag* (Perry, 2021). Tweets are annotated in context with all tweets in the same conversation path. We follow the taxonomy by Vidgen et al. (2021), but adapt the annotation scheme. We subsume identity-directed, affiliation-directed and person-directed abuse under abusive speech. The counterspeech category is left unchanged and refers to speech that "challenges, condemns or calls out the abusive language of others" (Vidgen et al., 2021). If an example could be annotated as both counterspeech and abusive speech, we label it as abusive if it is abusive on the same level as the preceding abusive tweet. Otherwise, e.g., if identity-directed abuse is countered with mild (!) person-directed abuse, we label it as counterspeech (cf. invicility vs. intolerance in Rossini, 2022). We extend the neutral language category to also include "non-hateful slurs" (Vidgen et al., 2021). We add a category "unintelligible" to explicitly mark examples which can not be assigned any meaningful label.

5.2 Results

Table 1 shows the bootstrap mean and confidence intervals of the absolute number of examples annotated as counterspeech for each approach. Again, all samples contain 810 examples. Both the multilingual model (85.54) and the keyword approach (81.73) achieve higher mean counts than the baseline (60.27). However, the lower ends of the confidence intervals of each method (model: 69.00, keywords: 65.00) lie within the baseline's confidence interval [46.00, 76.02]. Thus, improvements above the baseline are not statistically significant.

6 Discussion

Our results do not support the hypothesis that filtering a social media dataset using abusive keywords or a (multilingual) counterspeech classifier leads to a higher amount of true counterspeech.

[2] We merge the three evaluation samples to obtain a single annotated dataset (statistics in App. B). We release these annotations by the first author for further research: `https://github.com/morlikowski/german-counter-twitter`

	Mean	95% CI
Baseline	60.27	[46.00, 76.02]
Keyword	81.73	[65.00, 100.00]
Model	85.54	[69.00, 103.00]

Table 1: Mean and confidence intervals of the number of counterspeech examples based on bootstrap sampling each evaluation sample.

One reason might be that counterspeech is hard to detect, in particular when trying to distinguish it from abusive speech. In a quantitative analysis of the multilingual model's prediction errors, we find that among the falsely classified counterspeech comments, the fraction of ground truth abusive speech comments (20.00%) is slightly higher than in the entire sample (18.50%). We made similar observations in the results of the keyword approach, where a higher fraction of tweets predicted as abusive speech were actually counterspeech (14.40% vs. 10.40% in the entire sample). This suggests that abusive speech and counterspeech may share similar textual features to some extent. Indeed, in a qualitative analysis we find that both the correctly as well as the falsely classified counterspeech tweets, often contain harsh language against others, challenges of opinions or disagreements with opinions (see details in App. A.3).

We consciously decided to keep the counterspeech classifier comparatively simple, training on isolated comments. However, during annotation, we found that context is crucial to distinguish abusive speech from counterspeech. Thus, a multilingual model including context - preferably all preceding tweets in a conversation path - might help to increase the amount of counter speech examples. This importance of context is in line with findings by Yu et al. (2022) on the dataset we used for fine-tuning.

On the flip side, we can see our results as encouragement to not spend disproportional effort on improving filtering methods. The direct path to better counterspeech detection, not just for filtering, is by creating relevant resources. Our data collection method, derived from related work and exploratory analysis, seems to be able to produce fairly relevant samples even without additional filtering.

7 Conclusion

We asked whether filter methods can help to increase the amount of counter speech in a data sample when we do not have annotated data available for the given language. In a comparison of abusive keywords and a multilingual counterspeech classifier, both used to automatically assign labels for filtering, we do not find substantial evidence that they result in samples with more counterspeech.

Still, even without filtering, a purposive approach to data collection produces a meaningful sample with 7.4% true counterspeech in our setting. Thus, filtering datasets does not seem necessary for counterspeech annotation. Nevertheless, there is in principal a reasonable potential to improve filter methods. The proposed methods are necessarily imperfect, but better classifiers, e.g., trained on English Twitter data including context tweets, and improved keyword lists could change results. Also, a more extensive evaluation where multiple and larger samples are annotated might lead to different findings.

Acknowledgements

This work has in part been funded by the VolkswagenStiftung as part of the "Bots Building Bridges (3B)" project.

Limitations

All findings are limited to the German language and to the social media platform Twitter. Also, a crucial part of the study was conducted by only one person, in particular the evaluation annotation.

The results can only be related to the present data collection method and not to Twitter conversations in general. In the evaluation it could not be shown that the filter approaches find more counterspeech than the baseline. However, for the baseline we sampled from a specifically selected dataset resulting from seed accounts and filtering for conversation paths of specific lengths in conjunction with annotating root tweets. It is not clear whether the different approaches would work the same way on a differently collected dataset. The baseline sample already contains a fair amount of counterspeech. On a randomly selected dataset, the baseline sample might contain less counterspeech and the differences between the approaches could become more apparent.

Concerning the annotation, our results are also influenced by focusing on counterspeech in the direct conversational setting. In our definition, counterspeech can only be found in tweets that counter previous abusive speech in a conversation, not abusive speech in general. Sometimes comments were labelled as neutral although they countered an abusive point, because it referred to an expression of abusive language outside of Twitter and not in the preceding tweet.

In our evaluation, we decided to balance the requirement of in-context annotation, which needs complete conversation paths, and comparability of samples in a specific way: We make the samples comparable in the number of paths (200 in each sample), but as paths have varying lengths this inevitably results in a different number of tweets in each sample. Finding ways to meaningfully control for the number of tweets in each conversation path could lead to different results.

Ethical Considerations

We use the Twitter API to retrieve data for our study. Twitter can be considered a public space where discussions and posts are open to large audiences. Users have options to explicitly restrict the visibility of their posts in which case they are also not available via the API and are not used in our study. When users create a Twitter account, they give their consent for their (public) data to be shared with third parties[3]. In addition, for our initial data collection we selected accounts by people with a public profile who are often in public as part of their professional activities (e.g. politicians) and have high(er) follower counts. Users replying here have clear reason to assume public visibility for their posts, e.g., in contrast to tweets mentioning only users from a niche community. In addition, the dataset that is published with this study only contains IDs and not full tweets, so that the tweets' authors remain in control of their availability and can, for example, delete tweets at a later date which then can not be retrieved based on their IDs.

References

Abdullah Albanyan and Eduardo Blanco. 2022. Pinpointing fine-grained relationships between hateful tweets and replies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10418–10426.

Xiaoyu Bai, Flavio Merenda, Claudia Zaghi, Tomasso Caselli, and Malvina Nissim. 2018. Rug at germeval: Detecting offensive speech in german social media. In *Proceedings of the GermEval 2018 Workshop*, pages 63–70. Austrian Academy of Sciences.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. Washington, DC. United States Holocaust Memorial Museum.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. Countering online hate speech. UNESCO Publishing.

[3] https://help.twitter.com/rules-and-policies/data-processing-legal-bases

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Impact and dynamics of hate and counter speech online. 11(1):1–24.

Pierpaolo Goffredo, Valerio Basile, Bianca Cepollaro, and Viviana Patti. 2022. Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 57–66, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, page 90–94.

Rita Izsák. 2015. Report on the special rapporteur on minority issues. *Human Rights Council, Twenty-eight session*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on twitter. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, page 116–124.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on twitter. ArXiv:1812.02712.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):369–380.

Tal Perry. 2021. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rob Procter, Helena Webb, Marina Jirotka, Pete Burnap, William Housley, Adam Edwards, and Matthew L. Williams. 2019. A study of cyber hate on twitter with implications for social media governance strategies. In *Proceedings of the 1st Conference on Truth and Trust Online*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Patrícia Rossini. 2022. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425.

Julia Maria Struss, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *KONVENS*.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of* *the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Multilingual Model Approach

Model for Fine-tuning We use the multilingual pretrained Twitter model by Barbieri et al. (2022)[4]. The model is based on multilingual language model XLM-T (Conneau et al., 2020) which is additionally pretrained on millions of tweets in over thirty different languages. It fits to this work especially because it is familiar with social media language, German and English with which it will be fine-tuned to classify counterspeech.

Training Dataset There are many datasets available that contain hate- and counterspeech, yet most of them do not fit to the research as they do not contain real social media language but artificially generated data (Chung et al., 2019; Fanton et al., 2021), or they just contain the IDs of the tweets and not the text (Albanyan and Blanco, 2022), which would make it on the one hand harder to get the training data and on the other hand less efficient which Madukwe et al. (2020) showed. Another problem with some datasets (Fanton et al., 2021) was that they only contained hate- and counterspeech but no neutral speech which could, when training a model on it, lead to the model being overfitted and not recognising neutral speech.

That is why the data of Yu et al. (2022) seems to be the best fitting[5]. It contains real social media data from Reddit, is completely in English, the comments are available and do not have to be retrieved via the API and their annotation guidlines also use the definition of Vidgen et al. (2021), which means that their definition is close to the one this research is using. The dataset consists of what they call "parent" and "target" pairs. The parent always being a hateful comment the target being hate, counterhate or neutral. For the training of the model only the target comments were used and the context in which they were posted was ignored. The dataset consists of 6,846 parent-target pairs, with a label assigned to every target (hate, counter-hate, neutral). Yu et al. (2022) released with their work two datasets i) a gold dataset which consists of 4,751 pairs and ii) a silver dataset consisting of 2,095 pairs. The gold-set only includes data, with Krippendorff's alpha at or above 0.6, which estimates the inter-annotator agreement. Although the silver-set is noisy, they showed that it can be useful to

[4] https://github.com/cardiffnlp/xlm-t

[5] https://github.com/xinchenyu/counter_context

train models (Yu et al., 2022).

Preprocessing The preprocessing of the training data is in line with the procedure proposed of Barbieri et al. (2022). It is limited to replacing usernames (marked with an "@") with "@user" and hyperlinks with "http". As the training data ((Yu et al., 2022)) comes from Reddit, it does not contain usernames and also links were rare across the data.

Training For training the models we used the Huggingface Transformers library (4.30.2). In the code that Barbieri et al. (2022) make available on their github page[6], they suggest some hyperparameters for the training. Except for the batch size which was reduced to 4 because of computing power, the values of the learning rate (2^{-5}) and the number of the epochs (1) served as a basis for the finetuning of the model. Other than that we use the default parameters. We use the maximum sequence length of 512, with truncation and padding to the maximum length.

Yu et al. (2022) also suggest blending silver and gold data to enhance the model's performance. The silver data is a dataset in which there was a lot of ambiguity between the annotators (Krippendorff's alpha < 0.6), which is why the data was not used in the final version. However, Yu et al. (2022) were able to show that the performance of the model improves when the two subsets are blended while training. For this research the model was pretrained with the silver dataset and then trained on the gold subset. In both cases the models were evaluated on the test set of the gold dataset.

We use the same splits for train, evaluation and test set as suggested by Yu et al. (2022). While Yu et al. (2022) make available what they call "parent-target"-pairs (the abusive post and the reply to it), we only use the "target"-comment train our classifier. For the gold dataset, the training set contains 3325 comments, the evaluation and test set both consisting of 713 comments. For the silver subset the training set contained 1675 comments and the evaluation set 419 comments.

After exploring different learning rates close to the reported one (2^{-5}), we found that 2^{-4} achieved the best results, so we trained all the models that rate. The models (both on the silver and gold datasets) were trained for one and two epochs. More epochs need access to substantial computational resources (two epochs still run in well under seven hours on a free Google Colab notebook). Also, spending less time/resources is in line with our scenario where we want a necessarily imperfect boost to data relevance without extensive additional effort.

Evaluation on the Original English Test Set The results presented in Table (5) were achieved by the models on the test split of the gold data. The majority baseline always predicts the neutral class. The rows below represent the results with different training settings: training with the silver dataset, the gold dataset or pretraining with the silver dataset and then training with the gold dataset. The model was trained either one or two epochs on the data respectively. The models presented in the Table show the best achieving ones.

First one can observe that every model, regardless whether it was trained on the silver or on the gold dataset, performs much better than the majority baseline (weighted avg. F1-score 0.50 vs. 0.34). The second important observation is that in almost all cases, the models that were pretrained on the silver dataset and then trained on the gold dataset, outperform those models which were only trained on one of the data subsets, this finding agrees with the one Yu et al. (2022) have made.

The best model achieves an average F1 score of 0.59, and for the hate- and counterspeech classes it achieves F1 scores of 0.49 and 0.44 respectively (0.71 for the neutral class). This means that the best model performs worse than the one of Yu et al. (2022). Yet the best performing model of Yu et al. (2022) was trained with the context of the post (parent-comment). Hence comparing the scores of the here presented model with the one of Yu et al. (2022) trained in a similar way (only target data and no additional data of related task), the here presented one achieves a slightly worse weighted average F1 (0.59 vs. 0.61), a much worse F1 regarding the hate class (0.49 vs. 0.57), but a slightly better F1 for the counterspeech class (0.44 vs. 0.43) which is especially important for the underlying research question.

Evaluation: German Annotated Data Set 8 shows the results of the multilingual language model compared to the majority baseline. The low precision as well as the high recall for the counterspeech class indicate that the model is overfitted to the counterspeech class, which is also evident when looking at the confusion matrix (9). The error

[6] https://github.com/cardiffnlp/xlm-t

for the "false negatives", i.e. the tweets that were labelled as abusive or neutral speech but were actually counterspeech, is rather small (28 as opposed to 77 true positives) and can therefore be neglected. More interesting is the distribution of errors among the "false positives" the tweets that were classified as counterspeech but are actually abusive or neutral speech because the error is rather big (289 false positives vs. 77 true positives).

An interesting aspect lies in examining the distribution of different classes based on the true label among the "false positives", the falsely marked counterspeech comments, compared with the distribution in the entire sample.

For falsely classified counterspeech in the model sample, abusive speech represents 20% while neutral speech amounts to 59%, while in the entire sample, abusive speech accounts for 18.5% and neutral speech constitutes 70.2%.

These observations highlight a greater resemblance between counterspeech and abusive speech compared to their association with neutral speech. They share might also share similar semantics, as evidenced by the differing proportions in the "false-positive" subset compared to the whole sample.

A.2 Keyword-Approach

Figure 2 shows the frequency of the 20 most frequent matches from the keyword list in the initial dataset. There are indeed words that one would expect and that are often used in a hateful context (i.e. "Arsch" [ass] , "Fresse" [mouth; shut up] or "Troll" [troll]). But there are also words which can be either used in a hateful manner but which are as well used to express strong feelings of agreement or dissent (i.e. "Bock" [ram], "Mist" [crap], "Schande" [disgrace]). Some words can have an abusive meaning only in specific contexts (i.e. "Autofahrer" [cardriver], "Gehirn" [brain], "Hirn" [brain], "Krankheit" [disease]).

Evaluation of the Keyword Approach on the German Annotated Data Set For the keyword approach 506 were marked as potentially hateful according to the list. Of the tweets marked, 151 were actually hateful, 282 were actually neutral speech and 73 were counterspeech. On the other hand, 159 tweets that were actually hateful were not containing any hatewords.

Also in the keyword approach we examined the proportions of the true labels among the falsely marked abusive comments. The amount of counterspeech comprises 14.4% while neutral speech accounts for 55.7%, compared to the overall sample where counterspeech represents 10.4% and neutral speech constitutes 72.6%.

This observation shows that counterspeech and abusive speech share certain keywords according to the keyword list, which makes them in that way more similar to each other than to the neutral class.

A.3 Qualitative Analysis

To get a better insight into where our model fails, we looked at the results again and analysed which particular linguistic features they contain. For this purpose, the data classified by the model was divided into the different "error categories" which were: "true positives", i.e. counterspeech that was classified as such, "false positives", i.e. tweets that were classified as counterspeech but whose ground truth is neutral or abusive speech, and finally "false negatives", i.e. all those tweets that were classified as neutral or abusive speech but are actually counterspeech (see Table 10, 11 and 12). Of the "false positives" 40 examples were looked at in total (20 of the ones whose ground truth was "neutral speech" and 20 of the ones whose ground truth was "abusive speech"). Same goes for the "false negatives" where 24 examples were considered (20 of the ones which were falsely classified as "neutral speech" and 4 that were falsely classified as "abusive speech"). Of the "true positives" 20 examples were analysed. The different features that the tweets contain vary from category to category because not every feature makes sense for every error type.

In order to determine the features according to which the tweets were analysed, it was first looked for peculiarities in the tweets and later also examined the features used in the qualitative analyses of other papers (Yu et al., 2022; Albanyan and Blanco, 2022). A tweet could show one, multiple or no feature of the features below.

Analysis of the true Positives Table 10 shows that the true positives contain much sarcasm, irony or a rhetoric question which seems to suggest that the model does not struggle with that type of speech. Another feature most correctly classified tweets showed, is an attack against the previous speaker or another person. Also 65% of the tweets challenged the opinion of the previous speaker or disagreed with it. The final feature that was observed and also the most subjective one is, whether

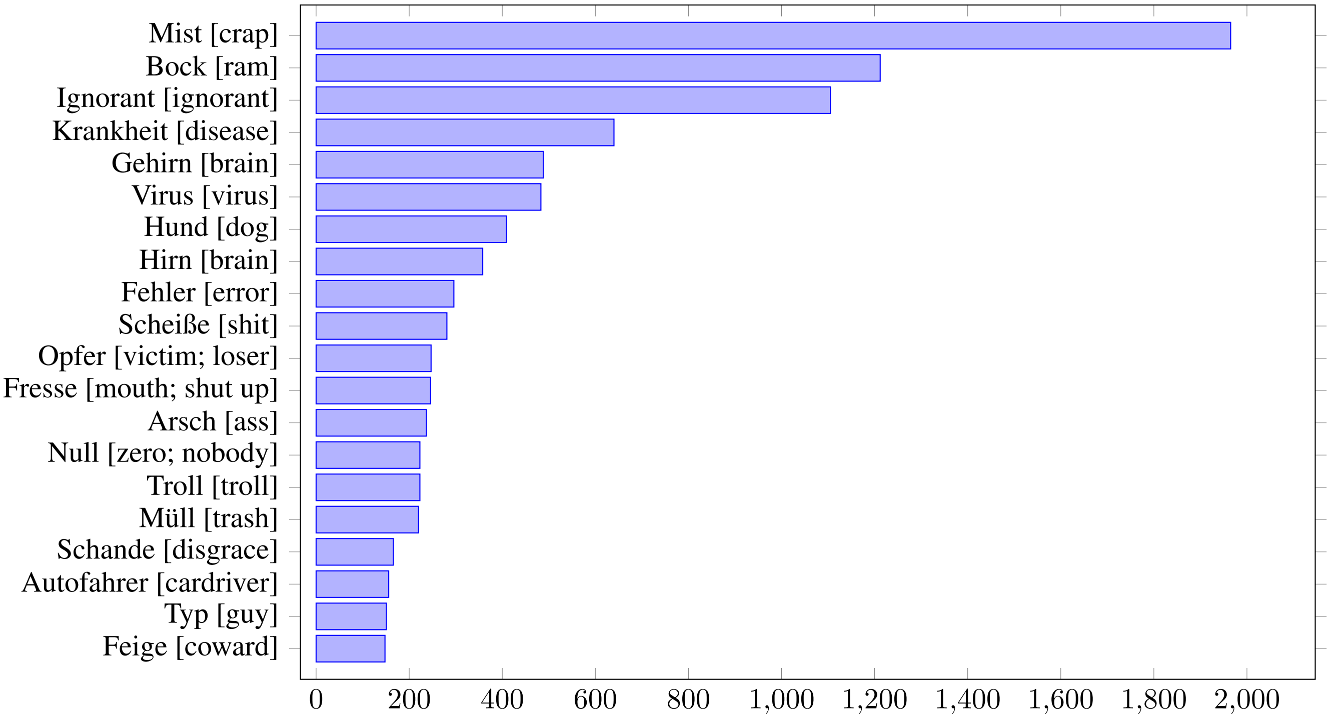

Figure 2: Frequency of the 20 most-frequently occurring keywords

a certain comment could be abusive speech in another context. It is noteworthy that 60% of the comments could also be labelled as abusive speech without knowing the context. That finding suggests that inaccuracies and misclassifications might especially happen between the counterspeech class and the abusive speech class which would also fit to the quantitative results that were shown in the confusion matrix of the multilingual language model.

Analysis of the false Positives Looking at the falsely positive classified comments (see Table 11), one notices that the amounts of irony (62.5 vs. 50), attacking a person (70 vs. 65) and challenging the opinion of the speaker or disagreeing with it (65 vs. 65), have similar values as in the qualitative analysis of the correctly classified counterspeech comments. 57.5% of the comments could have actually been classified as counterspeech by a human annotator, if they were read without or in another context. Although this category is rather subjective, it shows that counterspeech in its nature seems to contain features that are also contained in abusive speech and neutral speech (for example attacking a person, challenging an opinion or disagreeing with it). Also it shows that a lack of context makes it unclear even for a human annotator to classify a comment correctly, which agrees with the findings of Yu et al. (2022).

Analysis of the false Negatives The subset of the falsely negative classified counterspeech comments (see Table 12) makes up the smallest part of the errors (only 8.8% of the errors made regarding the counterspeech class were false negatives). The amount of irony is smaller than in the tables above but still rather high, a very low value is only adressed to the amount of rhetorical questions. That might but must not show that the model is inclined to classify comments that contain a rhetorical question as counterspeech. That would also fit to the observations made during the analysis of the falsely positive classified comments, which had a higher value in rhetorical questions. Attacks against people occur more seldomly and were especially present in those comments which were by the model labelled as abusive speech. Challenging the opinion or disagreeing with it also occurs less, as well as that the tweet could be hate in another context. These observations might show that the model is rather good in seeing disagreements but fails when counterspeech does not contain them and therefore classifiying it incorrectly.

Concluding Observations The quantitative analysis already showed that the model seems to be overfitted to the counterspeech class. Looking at the false positives it seems, that the model often classifies comments as counterspeech when they contain attacks against people or when they challenge or disagree with opinions, which apperently also happens in neutral and abusive speech a lot. Another big issue, concerning all comments, is the lack of context. Even for a human annotator the comment alone is often not sufficient to tell whether a comment is abusive, neutral or counterspeech. That is a valuable observation because it could be an indicator for further experiments in which the context could easily be included.

Another thing that is remarkable is that many falsely classified positives contain an attack of the speaker, same goes for the true positive class. That could be an indicator that the model is likely to classify comments, which are actually abusive speech, as counterspeech, because in ground true counterspeech comments, often the speaker is attacked in some kind of way.

B Tables and Figures

B.1 List of the Twitter-Accounts

Table 3 shows the list of accounts of public figures we identified which showed a high frequency of posting and which receive a lot of replies to their tweets, often containing abusive speech.

	Frequency in Dataset
Unique Tweets	2543
Counterspeech	297
Hatespeech	385
Neutral Speech	1748
Not Understandable	113
Roots	117
Paths	594
Speaker	1140

Table 2: Full statistics of the annotated dataset created from merging all three evaluation samples

Name	Twitter Name	Occupation
Karl Lauterbach	Karl_Lauterbach	Politician (SPD), Member of the Bundestag
Georgine Kellermann	GeorgineKellerm	Journalist, Transwoman
Ricarda Lang	Ricarda_Lang	Politician (Bündnis 90/Die Grünen), Member of the Bundestag
Annalena Baerbock	ABaerbock	Politician (Bündnis 90/ Die Grünen), Member of the Bundestag
Christina Lindner	c_lindner	Politician (FDP), Member of the Bundestag
Luisa Neubauer	Luisamneubauer	Climate Activist
Friedrich Merz	_FriedrichMerz	Politician (CDU), Member of the Bundestag
Sahra Wagenknecht	SWagenknecht	Politician (Die Linke), Member of the Bundestag
Beatrix von Storch	Beatrix_vStorch	Politician (AfD), Member of the Bundestag
Tino Chrupalla	Tino_Chrupalla	Politician (AfD), Member of the Bundestag
Julian Reichtelt	jreichelt	Journalist, Former editor-in-chief of the Bild

Table 3: List of People whose Twitter-accounts have been monitored

	Model			Keyword			Random		
interval	counter	abusive	neutral	counter	abusive	neutral	counter	abusive	neutral
lower	69.0	120.0	519.0	65.0	108.0	540.0	46.0	98.0	571.0
average	**85.54**	**140.21**	**543.568**	**81.726**	**127.561**	**566.257**	**60.236**	**115.713**	**596.041**
upper	103.0	161.0	570.0	100.0	147.02	592.0	76.02	136.0	619.0

Table 4: Mean and confidence intervals of the number of counterspeech, abusive speech and neutral speech examples across the three filter approaches based on 1500 bootstrap samples of size 810.

Model		Abusive			Counter			Neutral			Weighted Avg.		
ep. silver	ep. gold	P	R	F1	P	R	F1	P	R	F1	P	R	F1
-	-	0.00	0.00	0.00	0.00	0.00	0.00	0.51	1.00	0.67	0.26	0.51	0.34
1	-	0.46	0.52	0.49	0.83	0.06	0.11	0.60	0.82	0.69	0.62	0.56	0.50
2	-	0.56	0.49	0.52	0.49	0.39	0.43	0.65	0.75	0.69	**0.59**	**0.60**	**0.59**
-	1	0.58	0.48	0.53	0.35	0.04	0.07	**0.60**	**0.90**	**0.72**	0.53	0.59	0.51
-	2	0.60	0.47	0.53	0.41	0.42	0.42	0.67	0.74	0.70	**0.59**	**0.59**	**0.59**
1	1	0.52	0.57	0.54	0.65	0.19	0.30	0.62	0.80	0.70	0.60	0.60	0.56
1	2	0.48	0.60	0.54	**0.48**	**0.42**	**0.45**	0.70	0.64	0.69	0.59	0.58	0.58
2	1	**0.51**	**0.60**	**0.55**	0.56	0.16	0.25	0.63	0.79	0.70	0.58	0.59	0.56
2	2	0.57	0.43	0.49	**0.47**	**0.43**	**0.45**	0.65	0.77	0.71	**0.59**	**0.60**	**0.59**

Table 5: Training Results of the Model on the test Split of the gold Dataset

	Abusive			Neutral			Counter			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority Baseline	0.00	0.00	0.00	0.71	1.00	0.83	0.00	0.00	0.00	0.51	0.71	0.59
Keyword Approach	0.30	0.49	0.37	0.79	0.79	0.79	0.00	0.00	0.00	0.64	0.67	0.65

Table 6: Results of the keyword-approach on the annotated German dataset

		Predicted Label		
		Abusive	Neutral	Counter
True Label	Abusive	151	159	0
	Neutral	282	1036	0
	Counter	73	115	0

Table 7: Confusion matrix of the Keyword Approach on the annotated German dataset

	Abusive			Neutral			Counter			Weighted Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority Baseline	0.00	0.00	0.00	0.67	1.00	0.80	0.00	0.00	0.00	0.45	0.67	0.54
Model Approach	0.48	0.18	0.26	0.82	0.62	0.71	0.21	0.73	0.33	0.70	0.57	0.60

Table 8: Results of the classification of the multilingual model on the annotated German dataset

		Predicted Label		
		Abusive	Neutral	Counter
True Label	Abusive	30	68	73
	Neutral	29	406	216
	Counter	4	24	77

Table 9: Confusion matrix of the classification of the multilingual model on the annotated German dataset

Error Type	%	Example	Ground Truth	Predicted
Sarcasm or Irony	60	[...] Sie sind ein sehr taktvoller Mensch... [You are a very tactful person....]	Counter	Counter
Rhetorical Question	45	Und das hat jetzt mit dem Thema genau was zu tun? [And that has exactly what to do with the topic?]	Counter	Counter
Attacks Speaker or other Person	65	Sie reden so einen Mist [...] [You talk such crap [...]]	Counter	Counter
Challenges Opinion or disagrees	65	Aha, interessant. Fragen sie jemanden, der zB bosnische Vorfahren hat, ob er denn Moslem ist? Oder fragen sie auch jeden einzelnen Deutschen, welcher Religion er angehört? Auch wenn er "so richtig deutsch " aussieht? [Aha, interesting. Do they ask someone who has Bosnian ancestors, for example, whether he is a Muslim? Or do you also ask every single German which religion he belongs to? Even if he looks "really German"?]	Counter	Counter
Could be abusive in another Context	60	Wer keine Grammatik und Rechtschreibung beherrscht und zudem Unsinn trötet, sollte nicht posten. [Anyone who does not know grammar and spelling and who also toots nonsense should not post.]	Counter	Counter

Table 10: Frequencies of the different textual features of the correctly classified counterspeech (true positives)

Error Type	%	Example	Ground Truth	Predicted
Sarcasm and Irony	62.5	[...] kannst ja von einem Berg springen, wenn du nicht mehr leben willst. [You can jump off a mountain if you no longer want to live.]	Abusive	Counter
Rhetorical Question	37.5	Was tun Sie, um russische Angriffe zu verhindern? Also außer twittern? [What do you do to prevent Russian attacks? So apart from tweeting?]	Neutral	Counter
Attacks Speaker or other Person	70	[...] Typisch wenn man bis über beide Ohren in islamische Ärsche krabbelt. Man adaptiert das Verhalten dieses menschlichen Abfalls. [[...] Typical when you crawl up Islamic asses up to both ears. One adapts the behaviour of this human waste.]	Abusive	Counter
Challenges Opinion or disagrees	65	In DEU wird niemand unterdrückt.Wenn Sie sich so fühlen sollten, steht Ihnen der Rechtsweg offen. Oder die Politik, wenn Sie gestalten möchten, entsprechende Mehrheiten vorausgesetzt. [In GER, no one is oppressed. If you feel this way, the legal process is open to you. Or politics, if you want to shape it, provided you have the right majorities.]	Neutral	Counter
Could be Counter-speech in another Context	57.5	[...] ich Zweifel ein wenig an der Ernsthaftigkeit dieses Tweets. Corona können Sie nicht beenden. [I have a little doubt about the seriousness of this tweet. You can't end Corona.]	Neutral	Counter

Table 11: Frequencies of the different textual features among the falsely classified abusive or neutral speech (false positives)

Error Type	%	Example	Ground Truth	Predicted
Sarcasm or Irony	42	[...] Deine Vorfahren müssen echt mächtig stolz auf dich sein! [[...] Your ancestors must be very proud of you!]	Counter	Neutral
Rhetorical Question	8	[...] Deutsch am Bahnhof gelernt oder wie? [Learned German at the train station or what?]	Counter	Neutral
Attacks Speaker or other Person	37.5	[...] dann musst du ja wohl der Obertrottel sein wenn dir das Leid anderer Menschen am A... vorbei geht. [[...] Then you must be the biggest fool if you don't give a damn about other people's suffering.]	Counter	Abusive
Challenges Opinion or disagrees	45.8	Na dann machen wir hier mal einen kleinen Test. Von den drei Personen ist nur eine Person muslimischen Glaubens. Schauen wir mal, ob sie die Person erraten können. [Well, let's do a little test here. Of these three people, only one is Muslim. Let's see if you can guess who that person is.]	Counter	Neutral
Missing Context	37.5	Bitte einfach Zeitung lesen. Nachrichten hören. [Please just read the newspaper. Listen to the news.]	Counter	Neutral
Could be abusive in another Context	37.5	Kacktroll, sei still. [Shithead, shut up.]	Counter	Abusive

Table 12: Frequency of the different textual features among the falsely classified counterspeech (false negatives)

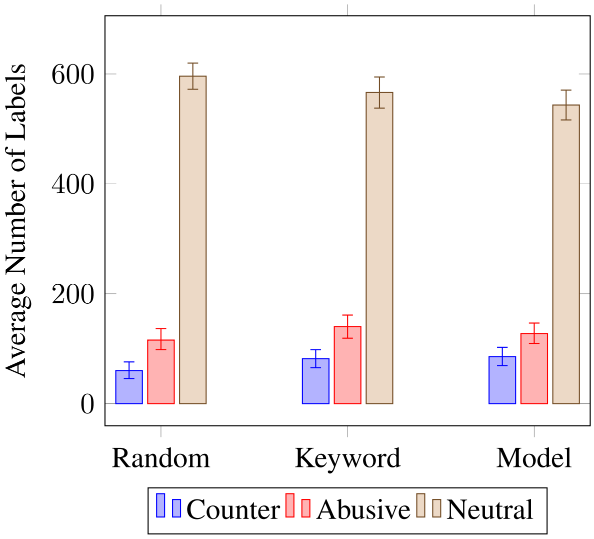

Figure 3: Mean and confidence intervals of the absolute amounts of counterspeech, abusive speech and neutral speech, across the three approaches, over 1500 bootstrap samples with size 810