# Data Integration for Toxic Comment Classification: Making More Than 40 Datasets Easily Accessible in One Unified Format

**Julian Risch** and **Philipp Schmidt** and **Ralf Krestel**
Hasso Plattner Institute, University of Potsdam
julian.risch@hpi.de, philipp.schmidt@student.hpi.de
ralf.krestel@hpi.de

## Abstract

With the rise of research on toxic comment classification, more and more annotated datasets have been released. The wide variety of the task (different languages, different labeling processes and schemes) has led to a large amount of heterogeneous datasets that can be used for training and testing very specific settings. Despite recent efforts to create web pages that provide an overview, most publications still use only a single dataset. They are not stored in one central database, they come in many different data formats and it is difficult to interpret their class labels and how to reuse these labels in other projects.

To overcome these issues, we present a collection of more than forty datasets in the form of a software tool that automatizes downloading and processing of the data and presents them in a unified data format that also offers a mapping of compatible class labels. Another advantage of that tool is that it gives an overview of properties of available datasets, such as different languages, platforms, and class labels to make it easier to select suitable training and test data.

## 1 Toxic Comment Datasets

Supervised machine learning and more specifically supervised deep learning is the current state-of-the-art for text classification in general and for toxic comment classification in particular (van Aken et al., 2018). The performance of these classifiers depends heavily on the size and quality of available training data, which is mostly used for fine-tuning general language models. The rather small sizes of annotated toxic comment datasets dates from the high costs for obtaining high-quality labels and the high variety of the task itself. For each language and each specific set of labels (racism, attack, hate, abuse, offense, etc.) new training and test datasets are needed. To circumvent this need, transfer learning can be adapted up to a certain degree (Bigoulaeva et al., 2021; Risch and Krestel, 2018). As a result, many researchers have created their own training and test datasets customized to their specific use cases. Three recent surveys compare and discuss datasets used in the literature for hate speech and abusive language detection (Madukwe et al., 2020; Poletto et al., 2020; Vidgen and Derczynski, 2020). These overviews help to assess the dataset landscape but stop short of doing the next step: integrating and unifying the various datasets and making them easily accessible.

In this paper, we present a software tool that provides easy access to many individual toxic comment datasets using a simple API. The datasets are in a unified data format and can be filtered based on metadata. The collection currently contains datasets in thirteen different languages: Arabic, Danish, English, French, German, Greek, Hindi, Indonesian, Italian, Marathi, Portuguese, Slovenian, and Turkish. Further, it covers a wide range of labels of different kinds of toxicity, e.g., sexism, aggression, and hate. The code is available in a GitHub repository[1] and also as a PyPI package[2] so that users can easily install it via the command *pip install toxic-comment-collection* and import datasets from the collection within python.

With our tool, researchers can combine different datasets for customized training and testing. Further, it fosters research on toxic comments and the development of robust systems for practical application. Important research and practical questions that can be investigated with our provided tool are:

1. How well do hate speech, toxicity, abusive and offensive language classification models *generalize across datasets*?

[1] https://github.com/julian-risch/toxic-comment-collection

[2] https://pypi.org/project/toxic-comment-collection

*Proceedings of the Fifth Workshop on Online Abuse and Harms*, pages 157–163
August 6, 2021. ©2021 Association for Computational Linguistics

2. What are the effects of different fine-tuning methods and *transfer learning*?

3. What is the relation of *different labeling schemes* and their effect on training?

4. Does toxic content look different on *different platforms* (Twitter, Wikipedia, Facebook, news comments)

5. How do *different language* influence classifier performance?

## 2 Unified Toxic Comment Collection

Creating a unified collection of toxic comment datasets comes with several challenges. First, the datasets are stored on various platforms and need to be retrieved. Second, different file formats of the datasets complicate data integration, and third, the different sets of class labels need to be mapped to a common namespace. This section describes how the creation of our collection addresses these two challenges and presents statistics of the collection.

### 2.1 Collection Creation

We consider all publicly accessible comment datasets for the collection that contain labels that are subclasses of toxicity, such as offensive language, abusive language, and aggression. The broad definition of toxicity as a higher-level concept builds a bridge between the different lower-level concepts. The term denotes comments that contain toxic language and was made popular by the Kaggle Challenge on Toxic Comment Classification in 2018, which defined toxic comments as comments that are likely to make a reader leave a discussion.[3] We exclude datasets that consider users instead of comments as the level of annotation (Chatzakou et al., 2017; Ribeiro et al., 2018) or study a different type of conversation, e.g., WhatsApp chats, where the participants presumably know each other in person (Sprugnoli et al., 2018).

The datasets that we collected come from various sources, such as GitHub repositories, web pages of universities, or google drive and other file storage platforms. Even more diverse than the different source platforms are the file formats of the datasets. From csv files with different column separators and quoting characters, over excel sheets, sql dumps, to txt files with single records spanning multiple rows, optionally compressed as zip or tar files — converting all these formats into the same standardized csv format of our collection is the second step of the data integration after the datasets are retrieved.

The third step focuses on the class labels. These labels are encoded in different ways. In the simplest format, there is a single column that contains one string per row, which is the class label. In some datasets, the class labels are encoded with integers, presumably to reduce file size. For multi-label classification datasets, the column might contain a list of strings or lists of integers. We unify the format of the labels to lists of strings.

More importantly, we create a mapping of class labels so that labels with the same meaning but different names are replaced with the same label. This mapping is stored in a configuration file and can be customized by users. Different use cases require different mappings. For example, one mapping can be used to map all datasets in the collection to a binary classification task of toxic and non-toxic comments. The next section describes the effect of this mapping on the toxic comment collection and other statistics of collection in the next section.

### 2.2 Collection Statistics

The collection contains comments in thirteen different languages, from twelve platforms, and with 162 distinct class labels (before mapping them to a smaller set of class labels). There is a large set of labels that occurs only in one dataset, with each label referring to a particular subclass of toxicity and target, e.g., female football players as in the dataset by Fortuna et al. (2019).

After combining similar names through our mapping strategy, 126 class labels remain, with 57 of them occurring in more than 100 samples. The total number of samples is currently 812,993. We are constantly adding more datasets.

As described in the previous section, a mapping can also be used to create a binary view on the collection with only two class labels: toxic and non-toxic. To this end, the class labels *none* (471,871 comments), *normal* (37,922 comments), *other* (2,248 comments), *positive* (4,038 comments), and *appropriate* (2,997 comments) are mapped to *non-toxic* (519,076 comments). The labels *idk/skip* (73 comments) are discarded and all other labels are mapped to *toxic* (293,844 comments).

Table 1 gives an overview of the collection by listing all datasets currently included in the collec-

[3] https://kaggle.com/c/jigsaw-toxic-comment-classification-challenge

tion together with their number of samples, source platform, language, and class labels. The table reveals that Twitter is the primary data source and that there is no common set of class labels. As per Twitter's content redistribution policy,[4] the tweets themselves were (in almost all cases) not released by the researchers but only the tweet ids. These ids allow re-collecting the dataset via the Twitter API. Our tool automatizes this process, which is also called re-hydration.

A challenge that is not visible in Table 1 is the inherent class imbalance of many datasets. For example, the class distribution of the dataset of attacking comments by Wulczyn et al. (2017) exhibits a bias towards "clean" comments (201,081 clean; 21,384 attack), whereas the dataset by Davidson et al. (2017) exhibits a bias towards "offensive" comments (19,190 offensive; 4,163 clean). The latter class distribution is not representative of the underlying data in general. It is due to biased sampling, similar to the issues that apply to the dataset by Zhang et al. (2018). Zhang et al. (2018) collected their dataset via the Twitter API by filtering for a list of keywords, e.g., *muslim, refugee, terrorist,* and *attack* or hashtags, such as *#banislam, #refugeesnotwelcome,* and *#DeportallMuslims.* This step introduces a strong bias because all hateful tweets in the created dataset contain at least one of the keywords or hashtags. Thus, the data is not a representative sample of all hateful tweets on Twitter, and models trained on that data might overfit to the list of keywords and hashtags. However, the advantage of this step is that it reduces the annotation effort: fewer annotations are required to create a larger set of hateful tweets. In fact, most comment platforms contain only a tiny percentage of toxic comments. Since research datasets are collected with a focus on toxic comments, they can be biased in a significant way. This focused data collection creates non-realistic evaluation scenarios and needs to be taken into account when deploying models trained on these datasets in real-world scenarios.

Figure 1 visualizes the overlap of the set of class labels used in the different datasets contained in the toxic comment collection. On the one hand, there are rarely any pairs of datasets with the exact same set of labels (yellow cells). Exceptions are datasets by the same authors. On the other hand, there are also only a few pairs of datasets with no common class label at all.

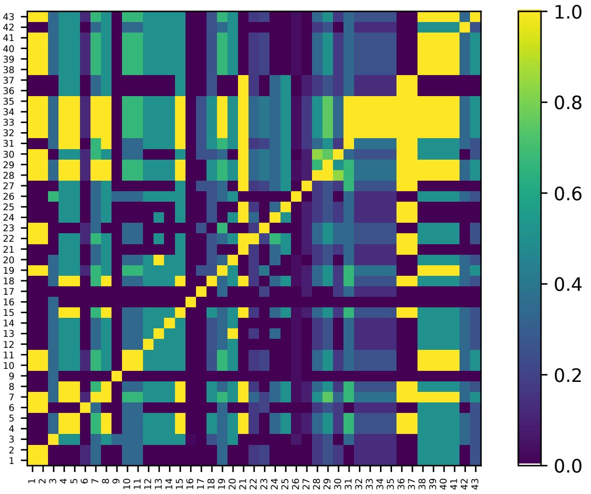

Figure 1: Heatmap of the pair-wise overlap of dataset class labels. Yellow cell color means that all class labels contained in the dataset of that row are also contained in the dataset of that column. See IDs in Table 1 for dataset names.

## 3 Conclusions and Future Work

In this paper, we addressed three challenges that hinder accessibility of research datasets of toxic comments: retrieving the datasets, unifying their file formats, and mapping their class labels to a common subset. To overcome these challenges, we present the toxic comment collection, which does not contain the datasets themselves, but code that automatically fetches these datasets from their source and transforms them into a unified format. Its advantages are the easy access to a large number of datasets and the option to filter by language, platform, and class label.

With the toxic comment collection, we aim to foster repeatability and reproducibility of research on toxic comments and to allow research on multilingual toxic comment classification by combining multiple datasets. We are continuously adding more datasets to the collection with routines to download them and to standardize their format automatically, e.g., we plan to integrate the datasets by Kumar et al. (2018) and Zampieri et al. (2019) next. We also plan to add contact information and instructions for datasets that are not publicly accessible but available only on request, such as the datasets by Golbeck et al. (2017), Rezvan et al. (2018), and Tulkens et al. (2016).

[4] https://developer.twitter.com/en/developer-terms/agreement-and-policy

Table 1: Datasets currently included in the toxic comment collection (sorted by year of publication). For this tabular presentation, we combined labels, e.g., *target* represents several different labels of targets.

| ID | Study | Size | Source | Lang. | Classes |
|---|---|---|---|---|---|
| 1 | Bretschneider and Peters (2016) | 1.8k | Forum | en | offense |
| 2 | Bretschneider and Peters (2016) | 1.2k | Forum | en | offense |
| 3 | Waseem and Hovy (2016) | 16.9k | Twitter | en | racism,sexism |
| 4 | Alfina et al. (2017) | 0.7k | Twitter | id | hate |
| 5 | Ross et al. (2016) | 0.5k | Twitter | de | hate |
| 6 | Bretschneider and Peters (2017) | 5.8k | Facebook | de | strong/weak offense,target |
| 7 | Davidson et al. (2017) | 25.0k | Twitter | en | hate,offense |
| 8 | Gao and Huang (2017) | 1.5k | news | en | hate |
| 9 | Jha and Mamidi (2017) | 10.0k | Twitter | en | benevolent/hostile sexism |
| 10 | Mubarak et al. (2017) | 31.7k | news | ar | obscene,offensive |
| 11 | Mubarak et al. (2017) | 1.1k | Twitter | ar | obscene,offensive |
| 12 | Wulczyn et al. (2017) | 115.9k | Wikipedia | en | attack |
| 13 | Wulczyn et al. (2017) | 115.9k | Wikipedia | en | aggressive |
| 14 | Wulczyn et al. (2017) | 160.0k | Wikipedia | en | toxic |
| 15 | Albadi et al. (2018) | 6.1k | Twitter | ar | hate |
| 16 | ElSherief et al. (2018) | 28.0k | Twitter | en | hate,target |
| 17 | Founta et al. (2018) | 80.0k | Twitter | en | six classes[d] |
| 18 | de Gibert et al. (2018) | 10.6k | Forum | en | hate |
| 19 | Ibrohim and Budi (2018) | 2.0k | Twitter | id | abuse,offense |
| 20 | Kumar et al. (2018) | 11.6k | Facebook | hing | aggressive |
| 21 | Mathur et al. (2018) | 3.2k | Twitter | en,hi | abuse,hate |
| 22 | Sanguinetti et al. (2018) | 6.9k | Twitter | it | five classes[b] |
| 23 | Wiegand et al. (2018) | 8.5k | Twitter | de | abuse,insult,profanity |
| 24 | Basile et al. (2019) | 19.6k | Twitter | en,es | aggression,hate,target |
| 25 | Chung et al. (2019) | 15.0k | misc | en,fr,it | hate,counter-narrative |
| 26 | Fortuna et al. (2019) | 5.7k | Twitter | pt | hate,target |
| 27 | Ibrohim and Budi (2019) | 13.2k | Twitter | id | abuse,strong/weak hate,target |
| 28 | Mandl et al. (2019) | 6.0k | Twitter | hi | hate,offense,profanity,target |
| 29 | Mandl et al. (2019) | 4.7k | Twitter | de | hate,offense,profanity,target |
| 30 | Mandl et al. (2019) | 7.0k | Twitter | en | hate,offense,profanity,target |
| 31 | Mulki et al. (2019) | 5.8k | Twitter | ar | abuse,hate |
| 32 | Ousidhoum et al. (2019) | 5.6k | Twitter | fr | abuse,hate,offense,target |
| 33 | Ousidhoum et al. (2019) | 5.6k | Twitter | en | abuse,hate,offense,target |
| 34 | Ousidhoum et al. (2019) | 4.0k | Twitter | en | abuse,hate,offense,target |
| 35 | Ousidhoum et al. (2019) | 3.3k | Twitter | ar | abuse,hate,offense,target |
| 36 | Qian et al. (2019) | 22.3k | Forum | en | hate |
| 37 | Qian et al. (2019) | 33.8k | Forum | en | hate |
| 38 | Zampieri et al. (2019) | 13.2k | Twitter | en | offense |
| 39 | Çöltekin (2020) | 36.0k | Twitter | tr | offense,target |
| 40 | Pitenis et al. (2020) | 4.8k | Twitter | el | offense |
| 41 | Sigurbergsson and Derczynski (2020) | 3.6k | misc | da | offense,target |
| 42 | Kulkarni et al. (2021) | 15.9k | Twitter | mr | negative |
| 43 | Kralj Novak et al. (2021) | 60.0k | Twitter | sl | offense,profanity,target,violent |

[a] argument,discrimination,feedback,inappropriate,sentiment,personal,off-topic

[b] aggression,hate,irony,offense,stereotype

[c] derailment,discredit,harassment,misogyny,stereotype,target

[d] abuse,aggression,cyberbullying,hate,offense,spam

## References

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. ACM.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval@NAACL)*, pages 54–63. ACL.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25. ACL.

Uwe Bretschneider and Ralf Peters. 2016. Detecting cyberbullying in online communities. In *Proceedings of the European Conference on Information Systems (ECIS)*, pages 1–14.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the Hawaii International Conference on System Sciences*, pages 2213–2222.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the International Web Science Conference (WebSci)*, page 13–22. ACM.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2819–2829. ACL.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 6174–6184. ELRA.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 512–515. AAAI Press.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, pages 11–20. ACL.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 42–51. AAAI Press.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, pages 94–104. ACL.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 491–500. AAAI Press.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 260–266. INCOMA Ltd.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the International Web Science Conference (WebSci)*, pages 229–233. ACM.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, pages 46–57. ACL.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data.

In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science (NLP+CSS@ACL)*, pages 7–16. ACL.

Petra Kralj Novak, Igor Mozetič, and Nikola Ljubešić. 2021. Slovenian twitter hate speech dataset IMSyPP-sl. Slovenian language resource repository CLARIN.SI.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. *arXiv preprint arXiv:2103.11408*.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1425–1431. ELRA.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, pages 150–161. ACL.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the Forum for Information Retrieval Evaluation*, page 14–17. ACM.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, pages 138–148. ACL.

Hamdy Mubarak, Darwish Kareem, and Magdy Walid. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, pages 52–56. ACL.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, pages 111–118. ACL.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684. ACL.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 5113–5119. ELRA.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

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 Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764. ACL.

Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the International Web Science Conference (WebSci)*, page 33–36. ACM.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgılio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 676–679. AAAI Press.

Julian Risch and Ralf Krestel. 2018. Aggression identification using deep learning and data augmentation. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@COLING)*, pages 150–158. ACL.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC@KONVENS)*, pages 6–9. University Frankfurt.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 2798–2805. ELRA.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3498–3508. ELRA.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, pages 51–59. ACL.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands Journal*, 6:3–20.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, pages 33–42. ACL.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12):1–32.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop@NAACL*, pages 88–93. ACL.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 1–10. Austrian Academy of Sciences.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 1391–1399. ACM.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1415–1420. ACL.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, pages 745–760. Springer.