Toxicity

Risch, Julian

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Risch, J. (2023). Toxicity. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 219-230). Berlin https://doi.org/10.48541/dcr.v12.13

Digital
Communication
Research.de

Recommended citation: Risch, J. (2023). Toxicity. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 219–230). Digital Communication Research.
https://doi.org/10.48541/dcr.v12.13

Abstract: In research on online comments on social media platforms, different terms are widely used to describe comments that are hateful or disrespectful and thereby poison a discussion. This chapter takes a theoretical perspective on the term toxicity and related research in the field of computer science. More specifically, it explains the usage of the term and why its exact interpretation depends on the platform in question. Further, the article discusses the advantages of toxicity over other terms and provides an overview of the available toxic comment datasets. Finally, it introduces the concept of engaging comments as the counterpart of toxic comments, leading to a task that is complementary to the prevention and removal of toxic comments: the fostering and highlighting of engaging comments.

DOI 10.48541/dcr.v12.13

Julian Risch

Toxicity

1 Toxic comments make readers leave a discussion

In computer science research in the broad field of social media analysis, *toxicity* is a collective term for a variety of phenomena. For several decades, comments have been denoted as toxic if they contain toxic language, including profanity, insults, and hate. As of today, there is no research in the computer science community that specifically addresses and discusses the term toxicity in this context. However, Fortuna et al. (2020) compared different terms across multiple datasets.

The term has become much more popular with the Kaggle Challenge on Toxic Comment Classification in 2018.[1] The general idea of such challenges or shared tasks is to stimulate research by having a competition, where all participants have access to the same training data, develop machine learning models in teams, and compare their model's performance with regard to a pre-defined machine learning task on the same test dataset and the same set of evaluation metrics. A machine learning task can be described as a set of given inputs, for example, social media comments, and expected outputs, for example, the two class labels *toxic* and *non-toxic*. A model that solves the task well can automatically map given inputs to the correct outputs, even for inputs it has not seen before. Common machine learning tasks besides

1 https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

classification are regression and clustering, where inputs are not mapped to previously defined class labels but to numeric values or to previously undefined groups of similar inputs. Kaggle is one of multiple web platforms where such tasks can be hosted as a competition and where the evaluation of the models is automated; for example, the number of correct predictions is calculated automatically.

Several series of shared tasks centered on toxic comment classification have recently emerged, with most of them having yearly or bi-yearly events. For example, HatEval deals with hate speech against immigrants and women (Basile et al., 2019), HaSpeeDe and HASOC with hate speech detection in general (Bosco et al., 2018; Mandl et al., 2019), IberEval assesses automatic misogyny identification (Fersini et al., 2018), GermEval and OffensEval cover offensive language (Struß et al., 2019; Zampieri et al., 2019), and TRAC focuses on aggression (Kumar et al., 2018; Bhattacharya et al., 2020). The main advantages of these competitions are the comparability of the results, the simultaneity of the efforts of the different teams, and thus, the intensive knowledge exchange at workshops typically following the competition. At these workshops, the final results are revealed, and the approaches are published in workshop proceedings.

This particular Kaggle Challenge on Toxic Comment Classification was organized by Google's subunit Jigsaw and allowed participants to compete at a shared task on a provided dataset. With more than 4,500 participating teams and a dataset of 150,000 hand-labeled comments, it was by far the largest shared task for toxic comment classification. The task defined toxic comments as comments that are likely to make a reader leave a discussion. Interestingly, this definition focuses on the effect of toxic comments on others instead of the linguistic features of the content.

The intention behind the definition becomes clearer with a closer look at the task's dataset, which is described in detail in a publication by Wulczyn et al. (2017). The dataset comprises comments that were posted on Wikipedia talk pages, where users discuss article page edits. Rude and disrespectful comments can arise in these discussions if users disagree on how an article page should be edited. In these situations, users might try to silence others and make users with other opinions leave the discussion to enforce their own views and end any further argument. The task definition of the Kaggle Challenge on Toxic Comment Classification includes not only a binary label for toxicity but also finer-grained labels: *toxic* and *severe toxic* as different severity levels of toxicity and a segmentation of toxic comments into

obscenity, threats, insults, and *identity hate* (each comment has been labeled by multiple crowd workers). Due to this fine-grained segmentation, the term *toxic* became established as a collective term for a variety of online comments.

What kind of content is considered toxic depends on the social media platform and its user community. Many platforms provide discussion guidelines that make transparent what rules users must adhere to and on what basis moderators remove content. However, in the end, it depends on the user community what kind of content makes them leave a discussion and is consequently considered toxic. Thus, the definition of toxicity depends on language use that is accepted by the community. For example, profanity might be allowed on some platforms and accepted by their users. While the wording of the definition leaves much room for interpretation, it is interpreted very similarly on many different platforms. The *netiquette,* the etiquette of the Internet, is a set of general guidelines that also applies to online discussions. For example, they are the basis for the discussion rules of online news platforms or Wikipedia.[2]

Due to its broad definition, the term toxicity can be applied to other comment datasets and platforms (van Aken et al., 2018), and this broadness can be seen as its main advantage. Other terms that are frequently used to denote hate speech in computer science research include offensive language, abusive language, and aggression. However, each of these terms describes only a subset of toxic comments. For example, vulgar or obscene language is not necessarily abusive, and benevolent sexism is not necessarily aggressive.[3] Toxicity as a higher-level concept, builds a bridge between the different lower-level concepts. As a consequence, models that need to be good at classifying a particular subset of toxic comments can be pre-trained on other similar subsets of toxic comments.

2 https://www.zeit.de/administratives/2010-03/netiquette/seite-2; https://www.welt.de/debatte/article13346147/Nutzungsregeln.html; https://de.wikipedia.org/wiki/Wikipedia:Wikiquette

3 Note that benevolent sexism is not necessarily perceived as benevolent by the recipient.

2 Toxic comment datasets

To show the diversity of toxicity and to give an overview of what falls under the definition of toxic comments, Table 1 lists the publicly available toxic comment datasets used in related work. While the term toxicity is rarely used in these datasets as a label, the labels used represent subclasses of toxicity. Most of the datasets have been labeled by the researchers themselves but a few of them by crowd workers. The respective publications contain descriptions of the individual datasets. Two recent surveys have compared and discussed the datasets (Poletto et al., 2020; Vidgen and Derczynski, 2020). A more detailed table that includes the number of comments per dataset and their language is also available (Risch, 2020). Table 1 makes clear that the majority of publicly available toxic comment datasets were collected on Twitter (26 out of 41). The set of class labels is more diverse. For example, there are datasets of comments from online news platforms where only one binary label is available, indicating whether the comment was published or removed from the platform (accept/reject). Further, there are different severity levels of toxicity (very toxic/mildly toxic), hate (strong hate/weak hate), and aggression (overtly aggressive/covertly aggressive). Many class labels focus on a particular subset of toxic comments, such as insults, profanity, cyberbullying, stereotypes, racism, and sexism.

Although the detection of toxic comments is challenging, the differentiation of subsets of toxicity is a difficult task on its own (van Aken et al., 2018). Davidson et al. (2017) and Kwok and Wang (2013) studied words that distinguish hate speech from offensive language. Some comments might fall into multiple subclasses, or they can happen to be at the borderline between two classes. An advantage of the term toxicity is that it does not require making a finer-grained and therefore more difficult classification. On the downside, the analysis of toxic comments is limited to a rather general level if no further fine-grained classification is used.

3 Toxic comments vs. engaging comments

Detecting and removing toxic comments prevents them from forcing readers to leave a discussion. Keeping more users engaged in an online discussion also matches the commercial interests of the providers of social media platforms.

They increase their revenue by maximizing the time that users spend on the platform. Thereby, they can show more ads and promote content to their users. An example is the shadow banning used by Twitter, where a toxic comment's visibility on the platform is reduced up to the point where it can only be seen by directly accessing the author's page.

An advantage of the term *toxic comment* over other terms is that it allows an elegant way of defining an opposite category of comments: while toxic comments make other users leave a discussion, engaging comments make other users join a discussion (Risch & Krestel, 2020).

The latter encourages users to actively join a discussion by replying to another user's comment or voting on a comment. Not only are these engaging comments thought-provoking, but they also stimulate users to express their opinions by posting a reaction. The concept of engaging comments has its roots in the concept of engaging, respectful and/or informative conversations (Napoles et al., 2017). A different definition considers constructiveness to be the opposite of toxicity (Kolhatkar & Taboada, 2017). However, constructiveness refers more to the content of the comment, whereas toxicity and engagement refer more to the effect of the comment. The two categories, toxicity and engagement, are not necessarily mutually exclusive. Comments can be rude and disrespectful, thereby making some users leave a discussion while at the same time, they can trigger some other users to join the discussion, either contributing counter-speech or, in the worst case, adding more toxic comments.

While social media platforms detect toxic comments to remove them, the detection of engaging (or constructive) comments would increase their visibility and highlight them on the platform. In the same direction, fostering engaging comments could help to have more diverse opinions in discussions, as it encourages more users to join a discussion.

4 Conclusion

Toxicity describes comments that make readers leave a discussion, for example, because of profanity, insults, threats, or hate speech. This chapter described the origins of this term and showed how it comprises the class labels used in various comment datasets. With this wide range being one main advantage of the term, the

Table 1: Toxic comment datasets (sorted by year of publication)

Study	Platform	Class labels
Kwok and Wang, 2013	Twitter	Racism
Djuric et al., 2015	News	Hate
Waseem, 2016	Twitter	Racism, sexism
Waseem and Hovy, 2016	Twitter	Racism, sexism
Badjatiya et al., 2017	Twitter	Racism, sexism
Davidson et al., 2017	Twitter	Hate, offense
Gao and Huang, 2017	News	Hate
Jha and Mamidi, 2017	Twitter	Benevolent/hostile sexism
Mubarak et al., 2017	News	Reject
Pavlopoulos et al., 2017	News	Reject
Schabus et al., 2017	News	Argument, discrimination, inappropriate, sentiment, off-topic
Vigna et al., 2017	Facebook	Strong/weak hate
Wulczyn et al., 2017	Wikipedia	Attack
Albadi et al., 2018	Twitter	Hate
Álvarez-Carmona et al., 2018	Twitter	Aggressive
Bosco et al., 2018	Facebook	Strong/weak hate
Bosco et al., 2018	Twitter	Aggression, hate, irony, offense, stereotype
Fersini et al., 2018	Twitter	Derailment, discredit, harassment, misogyny, stereotype, target
Founta et al., 2018	Twitter	Abuse, aggression, cyberbullying, hate, offense, spam
de Gibert et al., 2018	Forum	Hate
Kumar et al., 2018	Facebook	Aggression, covert, overt
Ljubešić et al., 2018	News	Reject

Sanguinetti et al., 2018	Twitter	Aggression, hate, irony, offense, stereotype
Wiegand et al., 2018	Twitter	Abuse, insult, profanity
Zhang et al., 2018	Twitter	Hate
Basile et al., 2019	Twitter	Aggression, hate, target
Fortuna et al., 2019	Twitter	Hate, target
Ibrohim and Budi, 2019	Twitter	Abuse, strong/weak hate, target
Kolhatkar et al., 2019	News	Very toxic, toxic, mildly toxic
Mandl et al., 2019	Twitter	Hate, offense, profanity, target
Mulki et al., 2019	Twitter	Abuse, hate
Ousidhoum et al., 2019	Twitter	Group, hostility, sentiment, target
Ptaszynski et al., 2019	Twitter	Cyberbullying, hate
Qian et al., 2019	Misc	Hate
Struß et al., 2019	Twitter	Abuse, insult, profanity, explicitness
Zampieri et al., 2019	Twitter	Offense, target
Bhattacharya et al., 2020	YouTube	Aggression, sexism, covert, overt
Caselli et al., 2020	Twitter	Abuse, explicitness
Çöltekin, 2020	Twitter	Offense, target
Pitenis et al., 2020	Twitter	Offense
Sigurbergsson and Derczynski, 2020	Misc	Offense, target

chapter also described the definition of its counterpart as another advantage. In contrast to toxic comments, engaging comments make readers join a discussion. Therefore, online platforms should detect both toxic comments and engaging comments to either increase or decrease their visibility. An interesting path for future work is to investigate the overlap of these two categories of comments.

Julian Risch is a senior machine learning engineer at deepset.

References

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 69–76.

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y- Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. *Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval@SEPLN)*, 74–96.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *Companion Proceedings of the International Conference on World Wide Web (WWW Companion)*, 759–760.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. *Proceedings of the International Workshop on Semantic Evaluation (SemEval@NAACL)*, 54–63.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., & Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@LREC)*, 158–168.

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the EVALITA 2018 hate speech detection task. *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, *2263*, 1–9.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 6193–6202.

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

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 512–515.

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, 11–20.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. *Companion Proceedings of the International Conference on World Wide Web (WWW Companion)*, 29–30.

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at IberEval 2018. *Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval@ SEPLN)*, 214–228.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 94–104.

Fortuna P., Soler, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? an empirical analysis of hate speech datasets. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 6786–6794.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 491–500.

Gao, L., & Huang, R. (2017). Detecting online hate speech using context aware models. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, 260–266.

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 46–57.

Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. *Proceedings of the Workshop on Natural Language Processing and Computational Social Science (NLP+CSS@ACL)*, 7–16.

Kolhatkar, V., & Taboada, M. (2017). Constructive language in news comments. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 11–17.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2019). The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, *4*(2), 155–190.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC@COLING)*, 1–11.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 1621–1622.

Ljubešić, N., Erjavec, T., & Fišer, D. (2018). Datasets of Slovene and Croatian moderated news comments. *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, 124–131.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in Indo-European languages. *Proceedings of the Forum for Information Retrieval Evaluation*, 14–17.

Mubarak, H., Kareem, D., & Walid, M. (2017). Abusive language detection on Arabic social media. *Proceedings of the Workshop on Abusive Language Online (ALW@ACL)*, 52–56.

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

Napoles, C., Tetreault, J., Pappu, A., Rosato, E., & Provenzale, B. (2017). Finding good conversations online: The yahoo news annotated comments corpus. *Proceedings of the Linguistic Annotation Workshop (LAW@EACL)*, 13–23.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4675–4684.

Pavlopoulos, J., Malakasiotis, P., Bakagianni, J., & Androutsopoulos, I. (2017). Improved abusive comment moderation with user embeddings. *Proceedings of the Natural Language Processing meets Journalism Workshop (NLPmJ@EMNLP)*, 51–55.

Pitenis, Z., Zampieri, M., & Ranasinghe, T. (2020). Offensive language identification in Greek. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 5113–5119.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation, 55*(1), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Ptaszynski, M., Pieciukiewicz, A., & Dybała, P. (2019). Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval Workshop*, 89–110.

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP- IJCNLP)*, 4755–4764.

Risch, J. (2020). *Reader comment analysis on online news platforms* (doctoral thesis). Universität Potsdam. https://doi.org/10.25932/publishup-48922

Risch, J., & Krestel, R. (2020). Top comment or flop comment? Predicting and explaining user engagement in online news discussions. *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 579–589.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2798–2805.

Schabus, D., Skowron, M., & Trapp, M. (2017). One million posts: A data set of German online discussions. *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, 1241–1244.

Sigurbergsson, G. I., & Derczynski, L. (2020). Offensive language and hate speech detection for Danish. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 3498–3508.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval task 2, 2019 shared task on the identification of offensive language. *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 352–363.

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. *Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP)*, 33–42.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS One*, *15*(12), 1–32.

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *Proceedings of the Italian Conference on Cybersecurity (ITASEC)*, 86–95.

Waseem, Z. (2016). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. *Proceedings of the Workshop on NLP and Computational Social Science (NLP+CSS@EMNLP)*, 138–142.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. *Proceedings of the Student Research Workshop@NAACL*, 88–93.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 1–10.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. *Proceedings of the International Conference on World Wide Web (WWW)*, 1391–1399.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings of the International Workshop on SemanticEvaluation (SemEval@NAACL)*, 75–86.

Zhang, Z., Robinson, D., & Tepper, J. A. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. *Proceedings of the Extended Semantic Web Conference (ESWC)*, 745–760.