**Dataset Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Type** | **Reference** | **Additional Notes:** |
| English Twitter Data Set | English Tweets annotated for hate speech; sexism/racism | Waseem and Hovy, NAACL 2016; Waseem, NLP and CSS 2016  <https://github.com/zeerakw/hatespeech> | 1,607 tweets, no context |
| German Twitter Data Set | Tweets regarding refugees in Germany; annotated with hate speech ratings; hate | Ross et al. NLP4CMC 2016  https://github.com/UCSM-DUE/IWG\_hatespeech\_public | 470 tweets, no context; could be used as foreign language comparison? |
| Wikipedia Abusive Language Data Set | Discussion comments from English Wikipedia; annotated on whether it contains personal attack + demographic data for crowd who annotated | Wikipedia Detox project  Wulczyn, Thain N., and Dixon L. (2017). “Ex machina: Personal attacks seen at scale”, in *ICWWW*, pp. 1391–1399.  <https://figshare.com/articles/Wikipedia_Detox_Data/4054689>  <https://arxiv.org/abs/1610.08914>  <https://meta.wikimedia.org/wiki/Research:Detox/Data_Release>  <https://github.com/ewulczyn/wiki-detox/blob/master/src/figshare/Wikipedia%20Talk%20Data%20-%20Getting%20Started.ipynb>  <https://conversationai.github.io/research.html> | Over 100k comments, each with 10 annotations by 4000 annotators. |
| WikiConv | Mulitlingual corpus encompassing history of conversations on Wikipedia Talk | Hua, Danescu-Niculescu-Mizil C., Taraborelli D., Thain N., Sorensen J., and Dixon L., (2018). “Wikiconv: A corpus of the complete conversational history of a large online collaborative community”, *arXiv preprint.*  <https://arxiv.org/pdf/1810.13181.pdf>  <https://convokit.cornell.edu/documentation/wikiconv.html>  <https://github.com/conversationai/wikidetox/tree/master/wikiconv> | includes deletion, modification, restoration of comments |
| Toxicity Detection w/ and w/o context | 2 datasets based on Wikipedia Talk pages; include context (context = parent comment + title of thread) | Pavlopoulos J., Sorensen J., Dixon L., Thain N., and Androutsopoulos I. (2020). “Toxicity Detection: Does Context Really Matter?”, in *Proc. of 58th Annual Meeting of Association for Computational Linguistics,* pp. 4296-4305.  <https://github.com/ipavlopoulos/context_toxicity>  (includes classifier code as well) | 1st has 250 comments (AB test, 2 groups of annotators – with/without context)  2nd has 20k comments (10k annotated with context, rest without)  Unbalanced – toxic comments rare |
| Civil Comments Toxicity Kaggle (CCTK) | English comments annotated for toxicity, subtypes and mentions of identities | <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification>  <https://conversationai.github.io/research.html> | 2M comments, no context, can evaluate unintended bias, 450,000 comments annotated with identities |
| Wikipedia Toxicity Kaggle | Crowdsourced dataset from English Wikipedia Talk pages with 4 toxicity subtypes | <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data>  <https://conversationai.github.io/research.html>  <https://meta.wikimedia.org/wiki/Research:Detox/Data_Release> | 160k/223,549? human labelled comments. 5000 annotators. No context |
| Wikipedia Machine Annotations of Talk Pages | Machine-labelled annotations of every English Wikipedia Talk page comments | <https://figshare.com/articles/Wikipedia_Talk_Corpus/4264973>  <https://conversationai.github.io/research.html>  <https://meta.wikimedia.org/wiki/Research:Detox/Data_Release> | All comments from 2001 to 2015, roughly 95M comments total.  (supports large scale analysis) |
| Davidson et al. (2017) | English Twitter; hate/Offensive comments. | T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In ICWSM, pages 512–515, Montreal, Canada | 24,783 tweets, no context. Collected using lexicon of hateful terms. |
| Zampieri et al. (2019a) | English Twitter; offensive comments | M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In NAACL. | 14,100 tweets, no context. |
| Gao and Huang (2017) | English Fox News article comments; hate | L. Gao and R. Huang. 2017. Detecting online hate speech using context aware models. In RANLP, pages 260–266. | 1,528 comments over 10 articles, title and preceding comments provided as context; annotations context-aware.  Small dataset, can’t reconstruct threads and assess parent comments, only 1 annotator |
| Wiegand et al. (2018) | German Twitter; insult/abuse/profanity | M. Wiegand, M. Siegel, and J. Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In Proceedings of GermEval. | 8,541 tweets, no context, could be used as foreign language comparison? |
| Pavolopoulos et al. (2017a) | Greek comments on Gazzetta.gr; rejection | J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. 2017a. Deep learning for user comment moderation. In 1st Workshop on Abusive Language Online, pages 25–35. | 1.6M comments, professional moderator decisions (context-aware) but no context in dataset, foreign language comparison?  Plain text comments + context not available |
| Mubarak et al. (2017) | Arabic comments on Aljazeera.net; obscene/offensive | H. Mubarak, K. Darwish, and W. Magdy. 2017. Abusive language detection on arabic social media. In 1st Abusive Language Workshop, pages 52–56, Vancouver, Canada. | 31,633 comments, provides title as context, annotators aware of context, foreign language comparison? |
| Synthetic Test Set | generated from templates using 50 identity terms; 50% toxic, 50% non-toxic across terms; to measure unintended bias; simple sentences | Borkan D., Dixon L., Sorensen J., Thain N., and Vasserman L. (2019). “Nuanced metrics for measuring unintended bias with real data for text classification”, in *Companion Proceedings of the 2019 World Wide Web Conference*,Association for Computing Machinery, pp. 491–500.  <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/66073ca7ac60ee38e93fc1d173a09cab65f2fef3.pdf>  Dixon L., Li J., Sorensen J., Thain N., and Vasserman L. (2018). “Measuring and Mitigating Unintended Bias in Text Classification”, in *Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society.* | 77k examples, not real-life scenarios, misses nuanced identity content, intentionally simple |
| Constructive Comments Corpus (C3) | English news comments from The Globe and Mail 2012-2016 annotated for constructiveness and toxicity (+ sub-characteristics of constructiveness) (got toxicity scores using Perspective) | Kolhatkar V., Thain N., Sorensen J., Dixon L., and Taboada M., (2020). “Classifying Constructive Comments”. *arXiv preprint arXiv:2004.05476*.  <https://www.kaggle.com/mtaboada/c3-constructive-comments-corpus>  <https://arxiv.org/pdf/2004.05476.pdf>  <https://researchdata.sfu.ca/islandora/object/sfu%3A2977>  <https://github.com/kvarada/constructiveness> | 12k comments, drawn from SOCC; (already been through some moderation), head comments (not replies), slightly higher in constructive comments than non-constructive (imbalanced), 10.3% of comments close to threshold (no consensus) |
| SFU Opinion and Comments Corpus (SOCC) | English news comments from The Globe and Mail 2012-2016; pairs articles and comments + reply structures and metadata | <https://github.com/sfu-discourse-lab/SOCC>  Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., Taboada, M., in press. The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. Corpus Pragmatics | 663k comments, 304k threads, 10k articles; (posted on website, already moderated) |
| Yahoo News Annotated Comments Corpus (YNACC) | news comments from Yahoo News (+1k from Internet Argument Corpus); capture sentiment, persuasiveness, tone; quality of threads – constructive, polite/aggressive | <https://github.com/cnap/ynacc>  <https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=83&guccounter=1>  Napoles C., Tetreault J., Pappu A., Rosato E., Provenzale B., (2017). “Finding good conversations online: The Yahoo News Annotated Comments Corpus”, in *Proceedings of the 11th Linguistic Annotation Workshop, EACL,* pp. 13–23. | 522k from 140k threads (9.2k comments coded at comment-level, 2.4k threads at thread-level), annotated by professional editors + untrained workers; constructiveness for threads not comments |
| SENSEI Social Media Annotated Corpus | Guardian news comments; constructiveness labels | <https://mailman.uib.no/public/corpora/2016-December/025781.html> | 1,845 comments from 18 articles |
| New York Times comments | news comments from New York Times in 2017 and 2018 | <https://www.kaggle.com/aashita/nyt-comments>  <https://developer.nytimes.com/docs/community-api-product/1/overview> | 2M comments, 9K articles |

**Notes:**

* Twitter datasets difficult to reuse as abusive tweets are removed by platform and textual content of tweet not available (can’t store outside platform) – so only have annotations and tweet IDs
* Most datasets in English, some in German/Greek/Arabic could be used for foreign language comparisons
* Most datasets do not give context or tell annotators about context
* No large toxicity dataset includes raw text of target and parent comments with links between them so can’t exploit conversational context
* Results vary between short and long comments – look at lengths of comments