

Title: Summary of datasets appendix for “The unappreciated role of intent in algorithmic moderation of abusive content on social media”

Authors: Xinyu Wang (1), Sai Koneru (1), Pranav Narayanan Venkit (1), Brett Frischmann (2), Sarah Rajtmajer (1)

Date: July 29th, 2025

Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix D: Summary of datasets

Table D3. *Summary of datasets for social media abuse.*

Reference	Source	Author-defined scope	Content type	Context provided	Target annotation	Intent mentioned
(Waseem & Hovy, 2016)	Twitter	Hate speech	Text	No	No	No
(Waseem, 2016)	Twitter	Hate speech	Text	No	No	No
(Golbeck et al., 2017)	Twitter	Online harassment	Text	No	No	Yes
(Chatzakou et al., 2017)	Twitter	Cyberbullying	Multimodal	Metadata	No	Yes
(Gao & Huang, 2017)	Fox News	Hate speech	Text	Conversation	No	No
(Davidson et al., 2017)	Twitter	Hate speech	Text	No	No	Yes
(Gao et al., 2017)	Twitter	Hate speech	Text	No	No	No
(Jha & Mamidi, 2017)	Twitter	Sexism	Text	No	No	No
(Van Hee et al., 2018)	ASKfm	Cyberbullying	Text	Conversation	Yes	Yes
(Fersini et al., 2018)	Twitter	Misogyny	Text	No	Yes	Yes
(Ribeiro et al., 2018)	Twitter	Hate speech	Multimodal	Metadata	No	No
(ElSherief et al., 2018)	Twitter	Hate speech	Text	No	Yes	No
(Founta et al., 2018)	Twitter	Abusive behavior	Multimodal	No	No	Yes
(Rezvan et al., 2018)	Twitter	Online harassment	Text	No	No	Yes
(Salminen et al., 2018)	YouTube, Facebook	Online Hate	Text	No	Yes	Yes
(M. Zampieri et al., 2019)	Twitter	Offensive language	Text	No	Yes	No
(J. Qian et al., 2019)	Reddit, Gab	Hate speech	Text	Conversation	No	No
(Ousidhoum et al., 2019)	Twitter	Hate speech	Text	No	Yes	No

Reference	Source	Author-defined scope	Content type	Context provided	Target annotation	Intent mentioned
(Basile et al., 2019)	Twitter	Hate speech	Text	No	Yes	No
(Mandl et al., 2019)	Twitter	Hate Speech, Offensive Content	Text	No	Yes	No
(Wijesiriwardene et al., 2020)	Twitter	Toxicity	Multimodal	Conversation	No	Yes
(Gomez et al., 2020)	Twitter	Hate speech	Multimodal	Metadata	No	No
(Vidgen et al., 2020)	Twitter	Hate speech	Text	No	No	Yes
(Caselli et al., 2020)	Twitter	Abusive language	Text	No	No	Yes
(Ziems et al., 2020)	Twitter	Cyberbullying	Text	Conversation	Yes	Yes
(C. J. Kennedy et al., 2020)	Twitter, Reddit, YouTube	Hate speech	Text	No	Yes	Yes
(Aggarwal et al., 2020)	Twitter, Reddit, Formspring	Cyberbullying	Text	No	No	No
(Suryawanshi et al., 2020)	Kaggle, Reddit, Facebook, Twitter, Instagram	Offensive language	Multimodal	Metadata	No	Yes
(Van Bruwaene et al., 2020)	Facebook, Instagram, Twitter, Pinterest, Tumblr, YouTube	Cyberbullying	Text	Metadata	No	Yes
(Grimminger & Klinger, 2021)	Twitter	Hate speech	Text	No	No	No
(Qureshi & Sabih, 2021)	Twitter	Hate speech	Text	No	No	No
(Salawu et al., 2021)	Twitter	Cyberbullying	Text	No	No	Yes
(Samory et al., 2021)	Twitter	Sexism	Text	No	No	No
(He et al., 2021)	Twitter	Hate speech	Multimodal	No	No	Yes
(Vidgen, Nguyen, et al., 2021)	Reddit	Abusive language	Text	Conversation	Yes	Yes

Reference	Source	Author-defined scope	Content type	Context provided	Target annotation	Intent mentioned
(Ashraf et al., 2021)	YouTube	Abusive language	Text	Conversation	No	Yes
(Mathew et al., 2021)	Twitter, Gab	Hate speech	Text	No	Yes	No
(Albanyan & Blanco, 2022)	Twitter	Hate speech	Text	Conversation	No	No
(Toraman et al., 2022)	Twitter	Hate speech	Text	No	No	No
(Thapa et al., 2022)	Twitter	Hate speech	Multimodal	Metadata	No	No
(B. Kennedy et al., 2022)	Gab	Hate speech	Text	No	Yes	Yes
(Albanyan et al., 2023)	Twitter	Hate speech	Text	Conversation	Yes	No