[6] Fundacion Barcelona Media (FBM). Caw 2.0 training datasets, 2009. http://caw2.barcelonamedia.org/.

[7] I. Gagliardone, D. Gal, T. Alves, and G. Martinez. *Countering online hate speech*. UNESCO Publishing, 2015.

[8] A. Halfaker. mwdiffs. https://github.com/mediawiki-utilities/python-mwdiffs.

[9] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[10] Impermium. Detecting insults in social commentary dataset, 2012. https://www.kaggle.com/c/detecting-insults-in-social-commentary.

[11] K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2004.

[12] K. Krippendorff. Reliability in content analysis. *Human communication research*, 30(3):411–433, 2004.

[13] I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

[14] M. J. Moore, T. Nakano, A. Enomoto, and T. Suda. Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, 28(3):861–867, 2012.

[15] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *WWW*, 2016.

[16] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[17] S. Pieschl, C. Kuhlmann, and T. Porsch. Beware of publicity! perceived distress of negative cyber incidents and implications for defining cyberbullying. *Journal of School Violence*, 14(1):111–132, 2015.

[18] B. Plank, D. Hovy, and A. Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*, pages 742–751, 2014.

[19] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths. A web of hate: Tackling hateful speech in online social spaces. In *TA-COS*, 2016.

[20] A. Schrock and D. Boyd. Problematic youth interaction online: Solicitation, harassment, and cyberbullying. *Computer-Mediated Communication in Personal Relationships*, pages 368–398, 2011.

[21] S. O. Sood, E. F. Churchill, and J. Antin. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2):270–285, 2012.

[22] N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter*, 13(2):50–64, 2012.

[23] Support and Safety Team. *Harassment Survey*. Wikimedia Foundation, 2015. https://upload.wikimedia.org/wikipedia/commons/5/52/Harassment_Survey_2015_-_Results_Report.pdf.

[24] J. R. Tetreault, E. Filatova, and M. Chodorow. Rethinking grammatical error annotation and evaluation with the amazon mechanical turk. In *NAACL-HLT*, 2010.

[25] R. S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287, 2010.

[26] M. A. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817, 2012.

[27] W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[28] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*, pages 88–93, 2016.

[29] D. Wiener. Negligent publication of statements posted on electronic bulletin boards: Is there any liability left after zeran. *Santa Clara L. Rev.*, 39:905, 1998.

[30] Wikimedia. Harassment consultation 2015. https://meta.wikimedia.org/wiki/Harassment_consultation_2015.

[31] Wikimedia. Machine-learning tool to reduce toxic talk page interactions. https://meta.wikimedia.org/wiki/2015_Community_Wishlist_Survey/Bots_and_gadgets\#Machine-learning_tool_to_reduce_toxic_talk_page_interactions.

[32] Wikipedia. Help:Talk pages. https://www.mediawiki.org/wiki/Help:Talk_pages.

[33] Wikipedia. Wikipedia:No personal attacks. https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks.

[34] Wikipedia. Wikipedia:Revision_deletion. https://en.wikipedia.org/wiki/Wikipedia:Revision_deletion.

[35] N. E. Willard. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research Press, 2007.

[36] E. Wulczyn, N. Thain, and L. Dixon. https://figshare.com/articles/Wikipedia_Detox_Data/4054689.

[37] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *CIKM*, 2012.

[38] J.-M. Xu, B. Burchfiel, X. Zhu, and A. Bellmore. An examination of regret in bullying tweets. In *HLT-NAACL*, pages 697–702, 2013.

[39] M. L. Ybarra and K. J. Mitchell. Youth engaging in online harassment: Associations with caregiver–child relationships, internet use, and personal characteristics. *Journal of adolescence*, 27(3):319–336, 2004.

[40] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In *WWW*, 2009.