**Literature Notes**

**Paper 1 – “Toxicity Detection: Does Context Really Matter?”**

* 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://www.aclweb.org/anthology/2020.acl-main.396.pdf
* Most datasets ignore context of posts, judging comments independently, without preceding comments – so systems ignore context when trained on dataset
* No other statistics on how often context affects perceived toxicity

**Paper 2 – “Nuanced metrics for measuring unintended bias with real data for text classification”**

* Borkan, 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.
* Toxicity = “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”

Other papers on reading list:

Hosseini, Kannan S., Zhang B., and Poovendran R. (2017). “Deceiving Google’s perspective api built for detecting toxic comments”, *arXiv preprint.*

Wulczyn, Thain N., and Dixon L. (2017). “Ex machina: Personal attacks seen at scale”, in *ICWWW*, pp. 1391–1399.

Kumar, Ojha A. K., Malmasi S., and Zampieri M. (2018). “Benchmarking aggression identification in social media”, in *TRAC*, Santa Fe, USA.

Zampieri, Malmasi S., Nakov P., Rosenthal S., Farra N., and Kumar R. (2019b). “Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)”, in *SemEval*.

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.*