World Academy of Science, Engineering and Technology
International Journal of Social and Business Sciences
Vol:15, No:12, 2021

Contextual Toxicity Detection with Data Augmentation

Authors : Julia Ive, Lucia Specia

Abstract : Understanding and detecting toxicity is an important problem to support safer human interactions online. Our work focuses on the important problem of contextual toxicity detection, where automated classifiers are tasked with determining whether a short textual segment (usually a sentence) is toxic within its conversational context. We use “toxicity” as an umbrella term to denote a number of variants commonly named in the literature, including hate, abuse, offence, among others. Detecting toxicity in context is a non-trivial problem and has been addressed by very few previous studies. These previous studies have analysed the influence of conversational context in human perception of toxicity in controlled experiments and concluded that humans rarely change their judgements in the presence of context. They have also evaluated contextual detection models based on state-of-the-art Deep Learning and Natural Language Processing (NLP) techniques. Counterintuitively, they reached the general conclusion that computational models tend to suffer performance degradation in the presence of context. We challenge these empirical observations by devising better contextual predictive models that also rely on NLP data augmentation techniques to create larger and better data. In our study, we start by further analysing the human perception of toxicity in conversational data (i.e., tweets), in the absence versus presence of context, in this case, previous tweets in the same conversational thread. We observed that the conclusions of previous work on human perception are mainly due to data issues: The contextual data available does not provide sufficient evidence that context is indeed important (even for humans). The data problem is common in current toxicity datasets: cases labelled as toxic are either obviously toxic (i.e., overt toxicity with swear, racist, etc. words), and thus context does is not needed for a decision, or are ambiguous, vague or unclear even in the presence of context; in addition, the data contains labeling inconsistencies. To address this problem, we propose to automatically generate contextual samples where toxicity is not obvious (i.e., covert cases) without context or where different contexts can lead to different toxicity judgements for the same tweet. We generate toxic and non-toxic utterances conditioned on the context or on target tweets using a range of techniques for controlled text generation(e.g., Generative Adversarial Networks and steering techniques). On the contextual detection models, we posit that their poor performance is due to limitations on both of the data they are trained on (same problems stated above) and the architectures they use, which are not able to leverage context in effective ways. To improve on that, we propose text classification architectures that take the hierarchy of conversational utterances into account. In experiments benchmarking ours against previous models on existing and automatically generated data, we show that both data and architectural choices are very important. Our model achieves substantial performance improvements as compared to the baselines that are non-contextual or contextual but agnostic of the conversation structure.

Keywords : contextual toxicity detection, data augmentation, hierarchical text classification models, natural language processing

Conference Title : ICCSS 2021 : International Conference on Computational Social Science

Conference Location : Amsterdam, Netherlands

Conference Dates : December 02-03, 2021

Open Science Index, Social and Business Sciences Vol:15, No:12, 2021 publications.waset.org/abstracts/142333.pdf