Aditi Chaudhari

Dr. Karen Mazidi

CS 4395.001—Human Language Technologies

8 April 2023

ACL Paper Summary

In their paper titled "From the Detection of Toxic Spans in Online Discussion to the Analysis of Toxic-to-Civil Transfer," the researchers John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffery Sorenson, and Ion Androutsopoulos study the task of toxic detection spans and then discuss how this identification can be useful in toxic-to-civil transfer. Pavlopoulos et al. define toxic content as "rude, disrespectful, or unreasonable posts that would make users want to leave the conversation" (3721). More specifically, toxic span is defined as the span of text in a post that makes the post toxic. For instance, if the post was "What if his opinion is that most other commenters are **idiots**? :-)" (with the toxic content being bolded), then the toxic spans would be {55, 56, 57, 58, 59, 60}, which signifies the position of each letter in the word 'idiots.'

In the past, toxicity detection systems have been created, yet these systems are trained on datasets that have been annotated on a post level to classify whether the post is toxic or not. Rather than assign toxicity labels to entire posts, Pavlopoulos et al. dive deeper to specifically detect toxic spans. Although previous research has focused on detecting spans in propaganda and hate speech, the problem of detecting toxic spans is much broader and encompasses more than just these categories. This is precisely why Pavlopoulos et al. decided to investigate this issue.

To construct a dataset for identifying toxic spans in a post, the researchers used crowd annotators to identify toxic spans within an existing dataset called Civil Comments. Following this, the inter-annotator agreement was calculated, and the ground truth of the dataset was established. The resulting dataset, named ToxicSpans, is made of 11,035 posts that have been annotated for toxic spans. After creating the dataset, the researchers considered two methods for identifying toxic spans. The first method involved using spaCy's Convolutional Neural Network (which is pre-trained for parsing, tagging, and entity recognition) in order to perform sequence labeling (tagging words). This model was named CNN-SEQ. The researchers also trained a bidirectional LSTM (BILSTM-SEQ), a fine-tune BERT (BERT-SEQ), and a SPAN-BERT (SPAN-BERT-SEQ) for toxic spans detection. The second method used a binary classifier to predict whether a post was toxic or not and then utilized attention as a rationale extraction mechanism at inference to attain toxic spans. The two classifiers that were used in this method were a BILSTM with deep self-attention (named BILSTEM+ARE) and BERT with a dense layer and sigmoid on the CLS embedding (named BERT+ARE).

The primary evaluation method used to determine the best performing models was the F1 score. The BILSTEM+ARE model (F1 score: 57.7%) performed much better than the BERT+ARE (F1 score: 49.1%) model. Interestingly, the BILSTEM+ARE (F1 score: 57.7%) model performed similarly to the BILSTEM-SEQ (F1 score: 58.9%) model despite the fact that the BILSTEM+ARE model was trained with toxic and non-toxic posts whereas the BILSTEM-SEQ was only trained with toxic spans annotations. The model that performed the best was the SPAN-BERT-SEQ (F1 score: 63.0%), which is pre-trained to predict spans.

Toxic spans detections are an important first step to solving the difficulties that exist with moderating content. For instance, human moderators, such as news portal moderators, often deal with lengthy comments and it can take them a long time to ensure that these comments are appropriate for a general audience. Detecting toxic spans through natural language processing is a huge first step to semi-automatic moderation and healthier online discussions!

Information on the Authors:

John Pavlopoulos is affiliated with the Department of Computer and Systems Sciences, Stockholm University, Sweden and the Department of Informatics, Athens University of Economics and Business, Greece. His work has been cited 3,570 times according to Google Scholar.

Leo Laugier is affiliated with the Télécom Paris, Institut Polytechnique de Paris, France. His work has been cited 117 times according to Google Scholar.

Alexandros Xenos is affiliated with the Department of Informatics, Athens University of Economics and Business, Greece. His work has been cited 13 times according to Google Scholar.

Jeffery Sorenson is affiliated with the technology company Google. His work has been cited 5,086 times according to Google Scholar.

Ion Androutsopoulos is affiliated with the Department of Informatics, Athens University of Economics and Business, Greece. His work has been cited 13,669 times according to Google Scholar. He is the author with the most citations.

Works Cited

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.