Quantifying Depressed Social Media During COVID-19: Information Retrieval With ML & NLP

Brent D. Davis*1, Dawn Estes McKnight2, Rumi Chunara3,4, Daniel J. Lizotte1,5, and Alona Fyshe2,6

1Department of Computer Science, Western University, London, ON
2Department of Computing Science, University of Alberta, Edmonton, AB
3Department of Computer Science & Engineering, New York University, New York, NY
4Department of Biostatistics, New York University, New York, NY
5Department of Epidemiology and Biostatistics, Western University, London, ON
6Department of Psychology, University of Alberta, Edmonton, AB, Canada

1 Abstract

The ongoing pandemic continues to disrupt the normal functioning of society in numerous ways, and symptoms of depression are on the rise [Ettman et al., 2020]. In this work, we explored how analysis of social media can reveal changes in the number of authors presenting depressive symptoms on social media using Twitter and Reddit.

We first assessed the level of depressive symptoms expressed in a large set of tweets. While there are some efforts for identifying depressive symptoms in tweets, they are limited in scope and typically do not account for contemporary online discourse surrounding the experience of depression [Hasan et al., 2014, Pedersen, 2015]. To ensure that our assessment accounted for contemporary discourse and symptoms and experience are a main topic of discussion. To further ensure that our assessment accounted for language that expresses depressive symptoms in a variety of contexts, rather than only when explicitly discussing the experience of depression, we also extracted all of the other Reddit posts of users who posted in /r/Depression. These user posts were extracted from all posts made by all authors in /r/Depression across all of Reddit for November and December 2019 (the most recent two months available in their entirety on Pushshift [Baumgartner et al., 2020]).

We then trained a GloVe word embedding on the posts made by users across Reddit who post in /r/Depression [Pennington et al., 2014]. Using the resulting word vectors, we then trained an author representation using the usr2vec [Amir and Wallace] method for both our /r/Depression authors and a sampled set of users to act as contrast against our archetypal example. This produces a high-dimensional representation of a user, based on a composite of the word representations we trained previously. Then, we used a linear kernel support vector machine (SVM) to find a separating hyperplane between these high dimensional representa-

*Corresponding author: bdavis56@uwo.ca
References


