Classifying Loneliness Expressions of Users in Twitter

By: Zehavit Ganon

Supervised by: Dr. Tomer Sagi

2021

Loneliness as a public health challenge was described by Guntuku et al.[1] *"Loneliness is a major public health epidemic and an estimated 17% of adults aged 18–70 in the USA have reported being lonely"*. Loneliness can be expressed online on social networks such as Twitter, Facebook, TikTok. Loneliness can be expressed in different ways: text, video, emojis. In general: loneliness and other emotions can be expressed by any way that the user can communicate via online platforms. Guntuku et al.[1] suggested a method for passive assessment that can lead to interventions targeted towards those who may be at risk of developing a severe mental health condition. With this purpose in mind, Guntuku et al.[1] built a language-based predictive model for loneliness. Natural-language processing was used to characterize the topics and diurnal patterns of users, posts and their association with linguistic markers of mental health. They obtained the age and gender estimates by using a lexica developed by Sap M, et al. [2]. The study has several limitations that we take into account in our proposed work as well. The study sample consists of social media users and is not representative of the general population. Posts mentioning the loneliness related words may have been metaphorical or partial. The effects presented in Guntuku et al.[1] method may not be specific to loneliness considering potential mental health. Guntuku et al.'s method can help to predict almost any medical question that we have a lexica and a social network data set for. The authors also mention that potentially they selected users with more posts than the average Twitter user. Guntuku et al.[1] evaluated their study by using a control group method. Counts of language features in the users with posts including the words lonely or alone were compared with the control group. In addition, by measuring area under curve (AUC) they evaluated if expressions of loneliness can be predicted in users' timelines. The predictive model was trained by using Random Forests on the training set and then evaluated on a test set to avoid overfitting.

In this study, we implemented Guntuku et al.'s[1] method on 616 million Tweeter messages. We parsed Guntuku et al.[1] lexica file with given classified terms and weighs, we built a predictive model using Pyspark operations then we applied the predictive model on a given data set. As a result we created two major outcomes: (1) a ranked file with presumably lonely users and (2) a ranked file with presumably lonely tweets. In addition to ranked users and tweets we calculated some descriptive statistics on the generated Pyspark data frames to obtain additional knowledge about the results.