Stress Analysis with Reddit

Fan Luo Ping-Hsun Lee
University of Arizona Columbia University

Abstract

College students face many challenges in their daily works. It's resulting in a huge impact on students' stress level and mental health. According to a report published by The Health Minds in 2020 [1], 37% of students suffer from any level of depression (severe, moderate, and slight), and 31% of students suffer from any level of anxiety. However, among these students with positive depression and anxiety screens, only 39% seek mental health consulting in the past 12 months. There are two reasons behind that: First, They are afraid of peer pressure. According to the same report, 51% of the respondents agreed that people will think less of someone who has received mental health treatment. Second, people might not be aware of his/her mental status and stress level.

In this project, we aim to extract information from student's social media activities to serve as an early warning for mental health issues. At first, we explored the comments and posts on piazza, as it is directly related to student's activities. However, the information we could find was more in summary statistics format that we could not extract more insights from. Hence, we chose to use reddit as the main data source instead in the end.

Reddit is known to be a widely used online forum and social media site among the college student demographic [2]. Due to its forum structure, we are able to obtain comments, posts, and information from certain communities, which is great for breaking down the problem by topic. Plus, thanks to prior researches [3], we directly obtained the dataset with the text data [3] collected from dedicated subreddit communities related to mental health.

In this project, we aim to achieve the following goals:

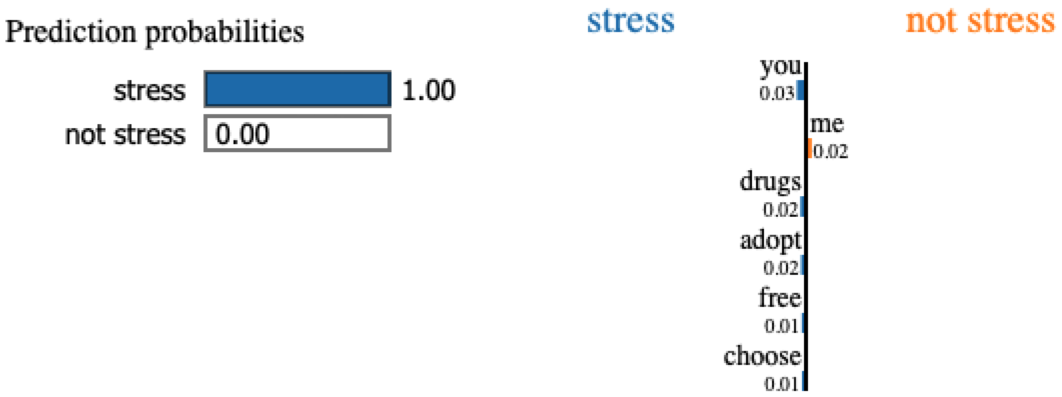
- **Aim 1:** Find the keywords list for the mental health issues, This could be saved for usage in future machine learning models as the word and id vectors.
- **Aim 2:** Construct a model that is able to prediction if the author of a post is facing stressful issues

Method

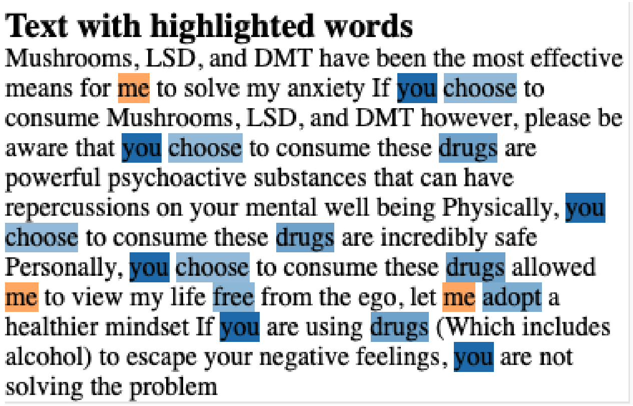
We use the dataset form Elsbeth and Kathleen [3]. We convert the raw text extract from reddit posts to a matrix of TF-IDF features with TfidfVectorizer of scikit-learn, and train a Logistic Regression model. To understand with terms contributed to the decision of the model, we apply the Lime to explain the judge of the classifier, in terms of detecting whether the author is stressful or not.

Result

For the performance of predicting stress, we get 0.812 in accuracy and 0.811 in F1. Below are the screenshots of lime explanations for one randomly selected sample.



Blue words indicate stress, while orange words indicate not stress.



Text with highlighted words

Mushrooms, LSD, and DMT have been the most effective means for me to solve my anxiety If you choose to consume Mushrooms, LSD, and DMT however, please be aware that you choose to consume these drugs are powerful psychoactive substances that can have repercussions on your mental well being Physically, you choose to consume these drugs are incredibly safe Personally, you choose to consume these drugs allowed me to view my life free from the ego, let me adopt a healthier mindset If you are using drugs (Which includes alcohol) to escape your negative feelings, you are not solving the problem

[1] Daniel Eisenberg, Peter Ceglarek, Sasha Zhou, "The Healthy Minds Study, 2020 Winter/Spring Data Report"

[2] Bagroy, Shrey, Ponnurangam Kumaraguru, and Munmun De Choudhury. "A social media based index of mental well-being in college campuses." Proceedings of the 2017 CHI Conference on Human factors in Computing Systems. 2017.

[3] Turcan, Elsbeth, and Kathleen McKeown. "Dreaddit: A Reddit dataset for stress analysis in social media." arXiv preprint arXiv:1911.00133 (2019).