A SENTIMENT AND TOPIC-BASED ANALYSIS OF MENTAL HEALTH DISCUSSION ON REDDIT

Joseph Tomasello and Daehan Kwak
Department of Computer Science and Technology, Kean University, Union, NJ 07083

Abstract

In the last decade, perceptions of mental health have undergone a significant transformation, shifting away from the stigma that once hindered open discussions. Today, conversations about mental well-being are prevalent, particularly online, where various platforms provide spaces for individuals to express their thoughts, share advice, and recount personal experiences. Reddit.com, in particular, has emerged as a prominent hub for such discussions, featuring numerous communities, or "subreddits," dedicated to mental health topics. Among the most popular are r/mentalhealth, r/depression, and r/Anxiety, which together have amassed over 1.8 million subscribers. This project aimed to analyze post data collected from these three subreddits over a four-year period, from March 11, 2018, to March 11, 2022. Utilizing natural language processing (NLP) techniques, the dataset of approximately 1.1 million posts was examined through both a sentiment analysis and the generation of a topic model. By exploring trends both collectively and within each individual subreddit, the study aimed to gain insights into how mental health discussions are approached by Reddit's user community and identify the most prevalent topics.

Introduction

- Across each of the 3 target subreddits, hundreds of users congregate daily to post about and discuss the topic of mental health from all conceivable angles
- Post data extracted from Reddit's leading mental health-centric communities can be used to provide a sense of how the subject is being examined by the site's userbase

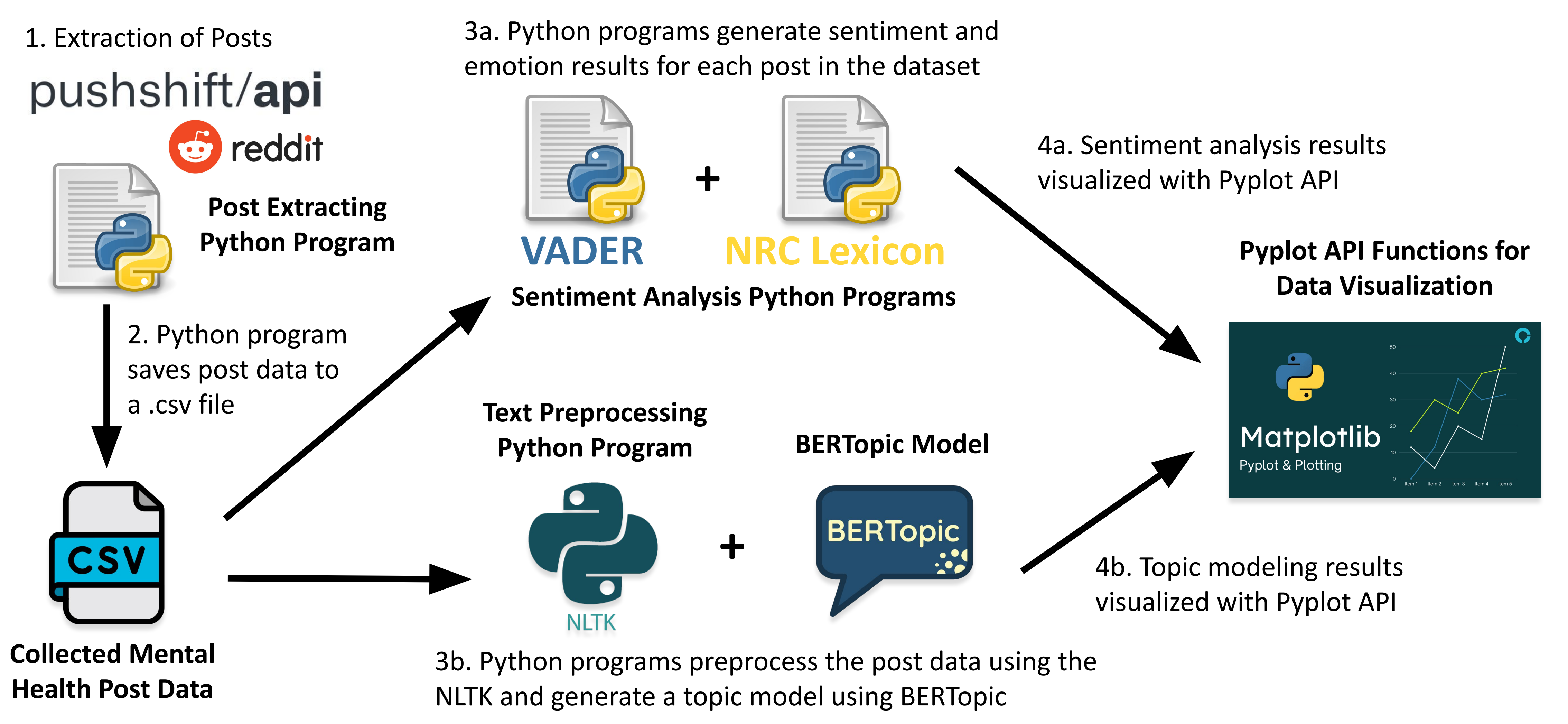
Objective

- Perform a sentiment analysis to gauge the predominant emotions expressed by users when discussing the topic of mental health on the site
- Produce a topic model to reveal the most prevalent post categories, along with any identifiable recurrent themes driving discussion of mental health on the site
- Visualize the results in these 2 areas in chart format

Methods / Implementation

- Access to Reddit and its archive of post data facilitated through use of the Pushshift API and saved to a .csv file.
- Target Subreddits: r/mentalhealth, r/depression, r/Anxiety
- Collection Period: March 11, 2018 – March 11, 2022
- Total Dataset Size: approx. 1.1 million posts
- Programming Language Used: Python
- Sentiment Analysis Tools
  - VADER: attuned to the identification of sentiments expressed on social media; capable of generating a score for each post indicative of an overall positive, negative, or neutral designation
  - NRC Emotion Lexicon: includes a range of 8 more pinpointed emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust
- Topic Modeling Tools
  - Natural Language Toolkit (NLTK): used to preprocess the textual post data
  - BERTopic Sentence Transformer Model (all-MiniLM-L6-v2)
- Graphing Tools
  - Pyplot API for Matplotlib

Approach to Extracting Posts and Subsequent Analysis


1. Extraction of Posts

Post Extracting Python Program

2. Python program saves post data to a .csv file

Collected Mental Health Post Data

3a. Python programs generate sentiment and emotion results for each post in the dataset

VADER + NRC Lexicon

Sentiment Analysis Python Programs

3b. Python programs preprocess the post data using the NLTK and generate a topic model using BERTopic

Text Preprocessing Python Program

BERTopic Model

4a. Sentiment analysis results visualized with Pyplot API

4b. Topic modeling results visualized with Pyplot API

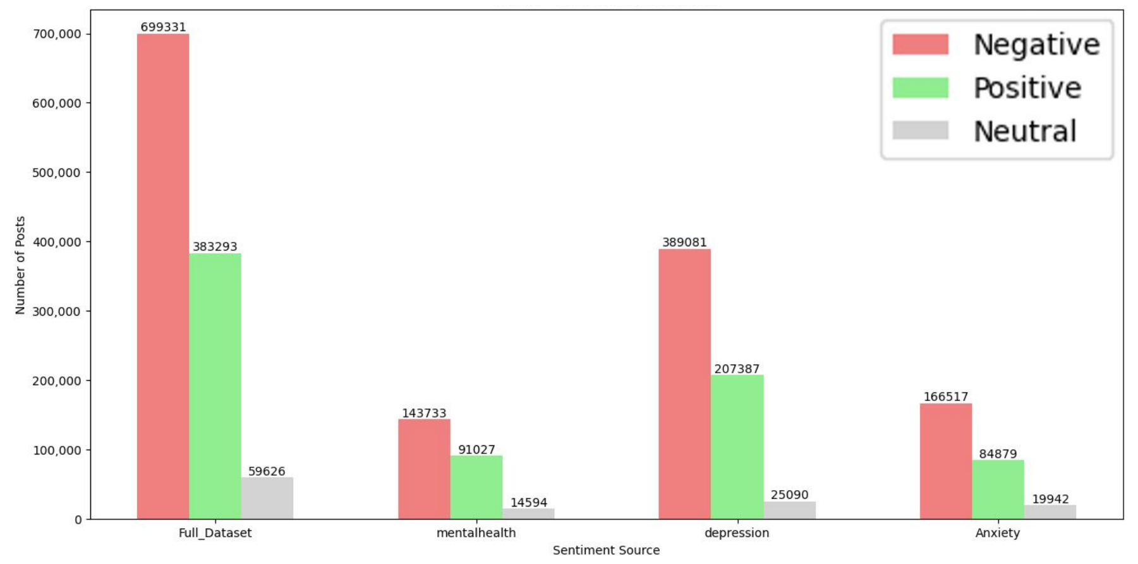
Pyplot API Functions for Data Visualization

Conclusion

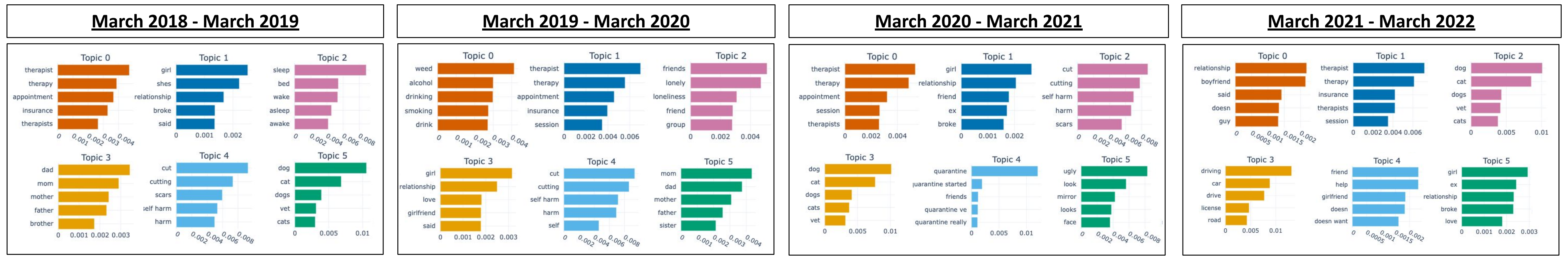
- VADER Sentiment Findings: indicated a majority of the dataset's posts could be categorized as negative in tone, with neutral posts accounting for a clear minority
- NRC Emotion Lexicon Findings: topped by the emotions sadness, fear, trust, and anticipation
  - Most of the 8 emotions remained consistent when expressed as a percentage of the overall emotion score total, with only fear, trust, and anticipation swapping positions over the course of the collection period
- Tag-Based Findings: pointed to "Question", "Need Support", and "Venting" as the most common flairs used on r/mentalhealth to categorize post content, with "Advice", "DAE Questions", and "Medication" being the most prevalent on r/Anxiety
- BERTopic Findings: the most frequently addressed topics were related to therapy, romantic relationships, pets, self harm, family, friendship, and substance abuse

Results

VADER – Sentiment Post Count Totals



BERTopic – Topic Word Scores per Year

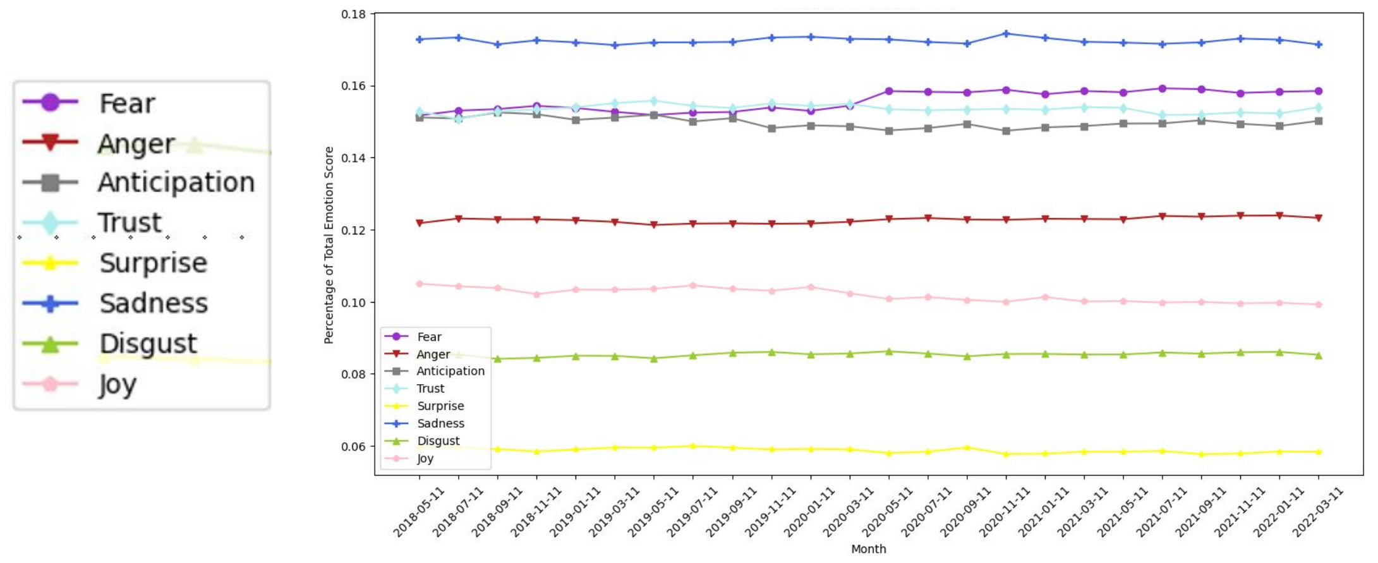

March 2018 - March 2019

March 2019 - March 2020

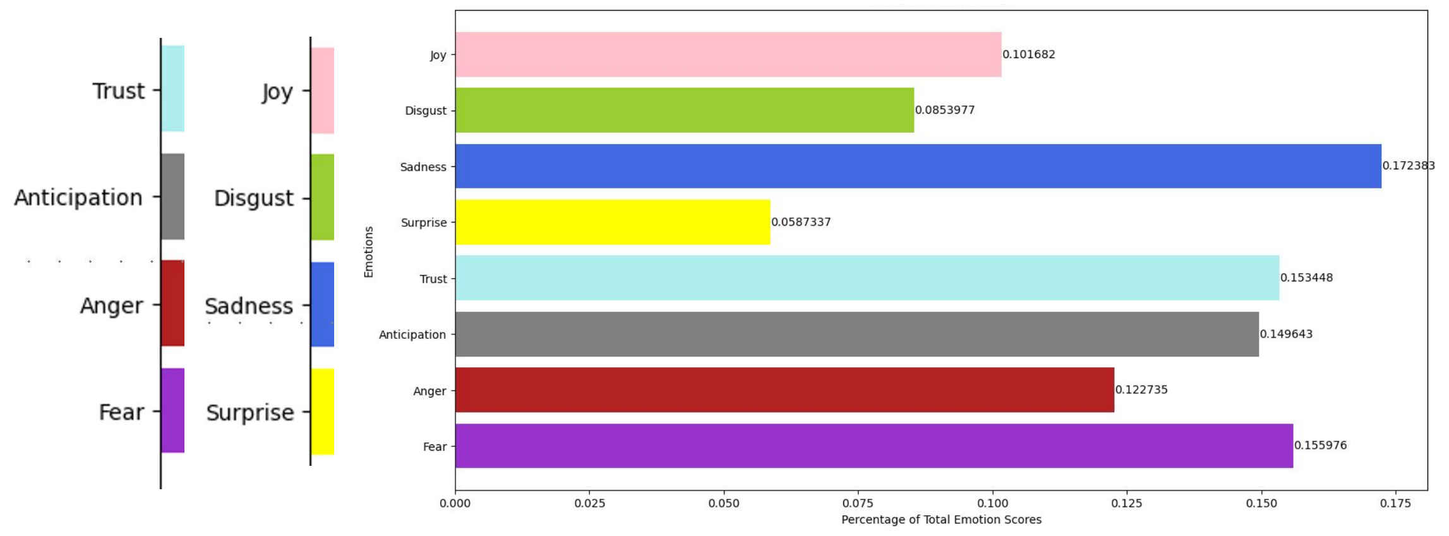
March 2020 - March 2021

March 2021 - March 2022

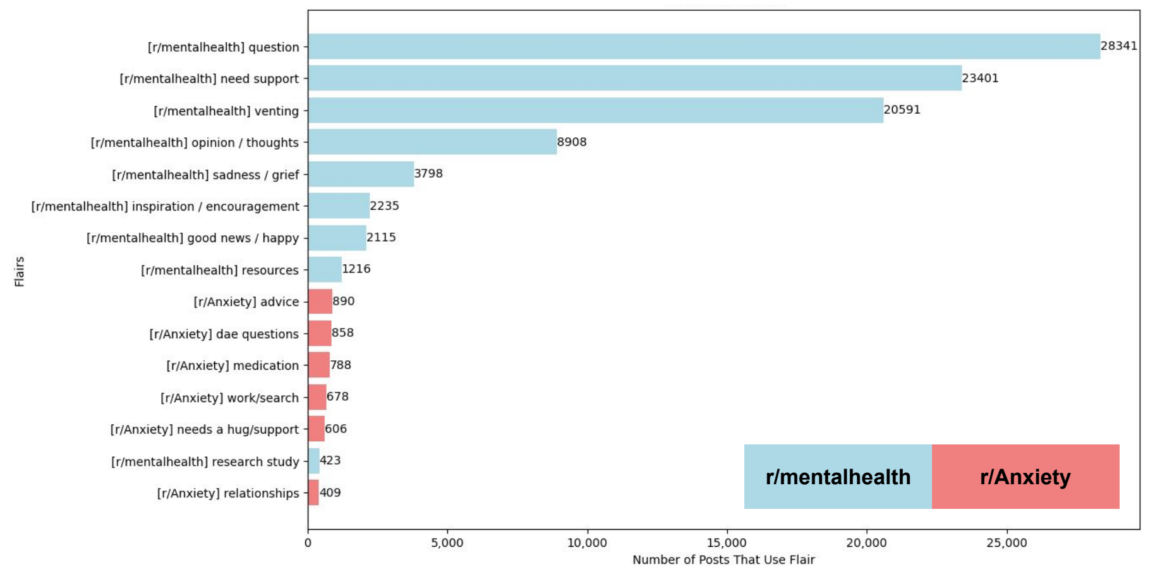
NRC Lexicon – Emotion Scores Over Time



NRC Lexicon – Emotion Score Totals



Flair Count Totals



Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants numbers CNS-2137791, and HRD-1834620. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References