REDDIT AS A FORUM FOR DISCUSSIONS ON MENTAL HEALTH

Joseph Tomasello and Daehan Kwak

School of Computer Science and Technology, Kean University, Union, NJ 07083

Abstract

Over the last decade, mental health has emerged as a topic of intense interest to the general populace, having finally shed much of stigmatization that kept it out of public discourse for so long. In fact, discussions on the subject are cropping up seemingly everywhere, particularly online, with many popular social-oriented platforms now featuring spaces set up specifically for users to offer up any relevant thoughts, opinions, advice, or even first-hand experience they may have with the subject. Some of the most heavily trafficked examples over the past few years can be found on Reddit.com, an American social news discussion platform divided into individual forums categorized by content, referred to as either “communities” or “subreddits” on the site. The 3 largest subreddits within the domain of mental health are the aptly-named r/mentalhealth, r/depression, and r/Anxiety, boasting over 1.8 million subscribers between them. This project set out to collate 4 years’ worth of post-based textual data from these 3 Reddit communities via the Pushshift API and then, using natural language processing (NLP) methodologies, anatomize the resulting dataset both in terms of a sentiment analysis and topic model.

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

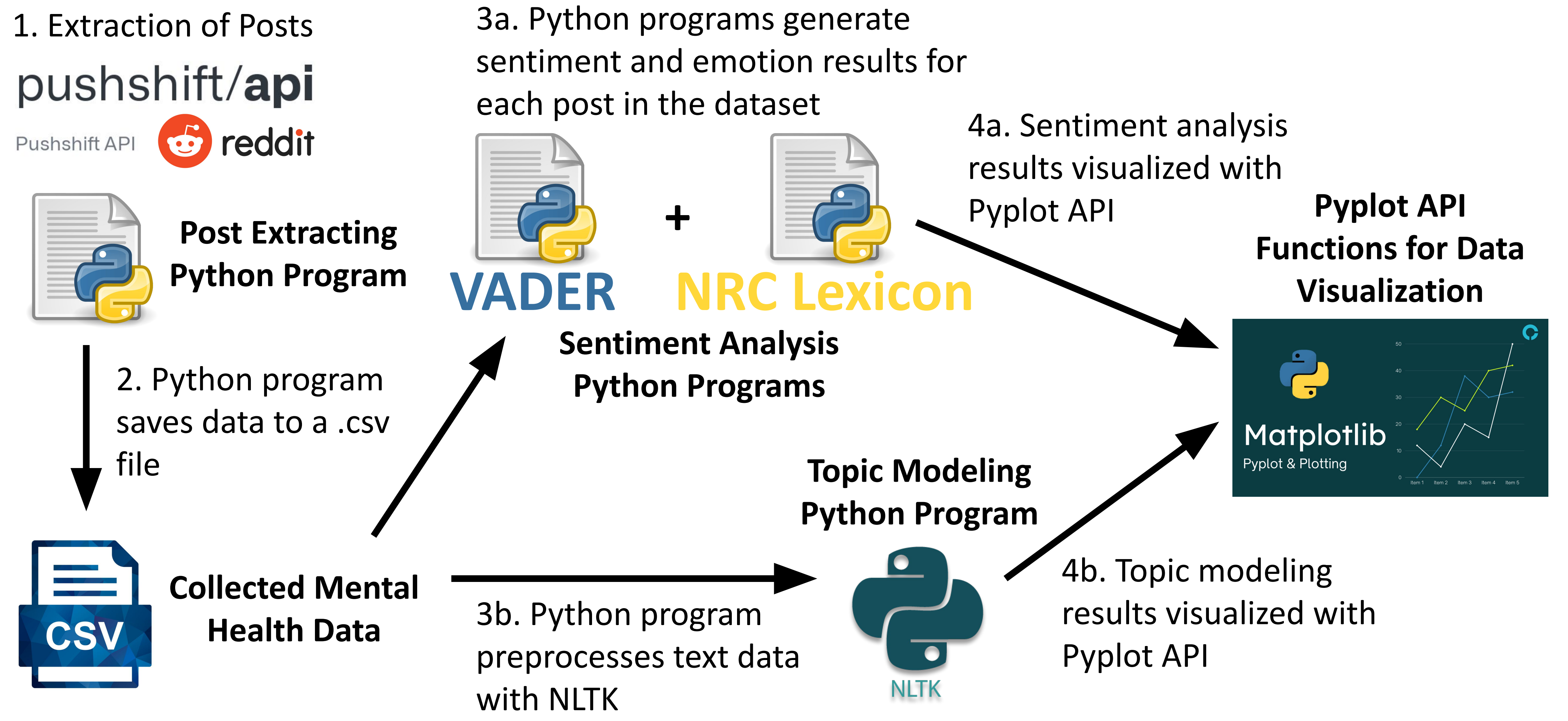
Objective

- Perform a sentiment analysis to gauge the emotional response of 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 of these 2 focus areas in a graphical 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. 1.25 million posts
- Programming Language Used: Python
- Sentiment Analysis Tools
  - VADER: attuned to the identification of sentiments expressed in 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 data
- Graphing Tools
  - Pyplot API for Matplotlib

Approach to Extracting Posts and Subsequent Analysis



1. Extraction of Posts

pushshift/api

Pushshift API reddit

Post Extracting Python Program

2. Python program saves data to a .csv file

Collected Mental Health Data

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

VADER + NRC Lexicon

Sentiment Analysis Python Programs

3b. Python program preprocesses text data with NLTK

Topic Modeling Python Program

4a. Sentiment analysis results visualized with Pyplot API

4b. Topic modeling results visualized with Pyplot API

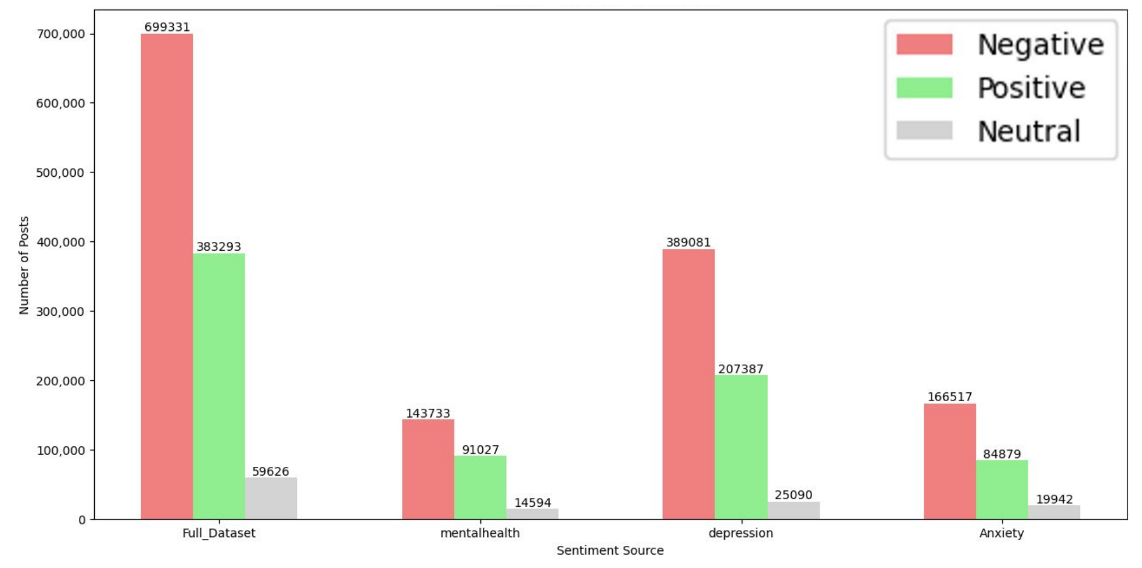
Pyplot API Functions for Data Visualization

Conclusion

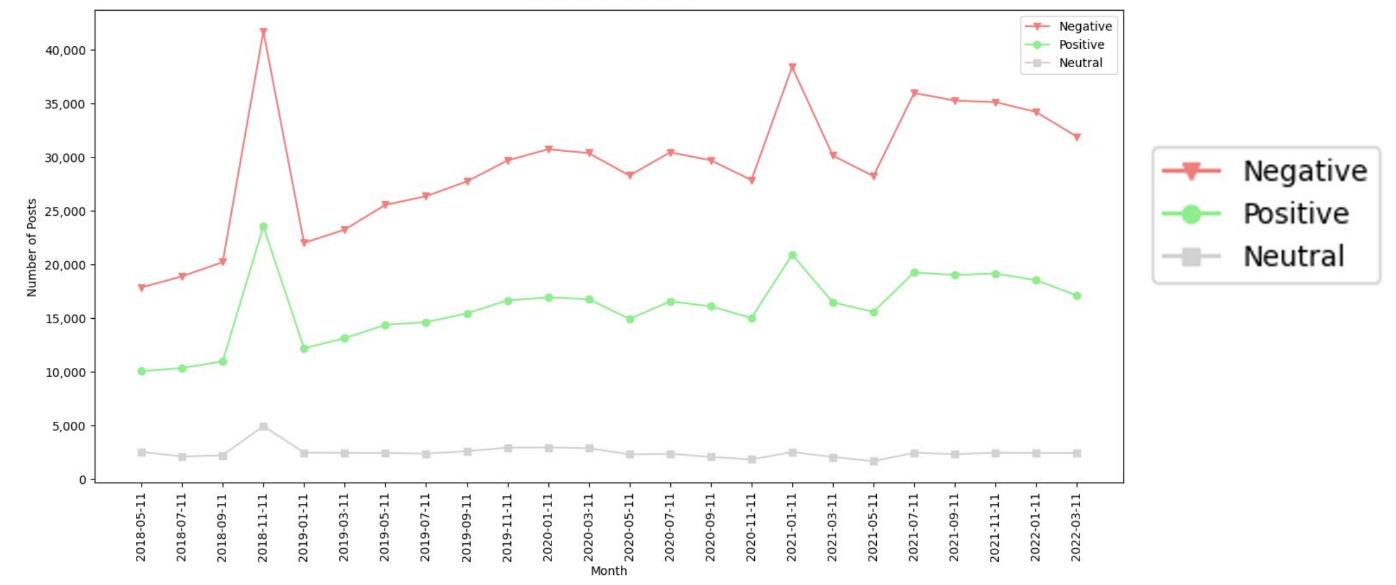
- VADER Sentiment Findings: indicated a majority of the dataset’s posts could be categorized as negative in tone, with neutral posts accounting for a clear minority
- NRC Emotion Lexicon Findings: topped by the emotions sadness, fear, trust, and anticipation
  - Most of the 8 identifiable emotions remained 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
- Topic Modeling Results: 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 over on r/Anxiety
- Top 5 Terms in Dataset: “time”, “life”, “people”, “friend”, and “anxiety”

Results

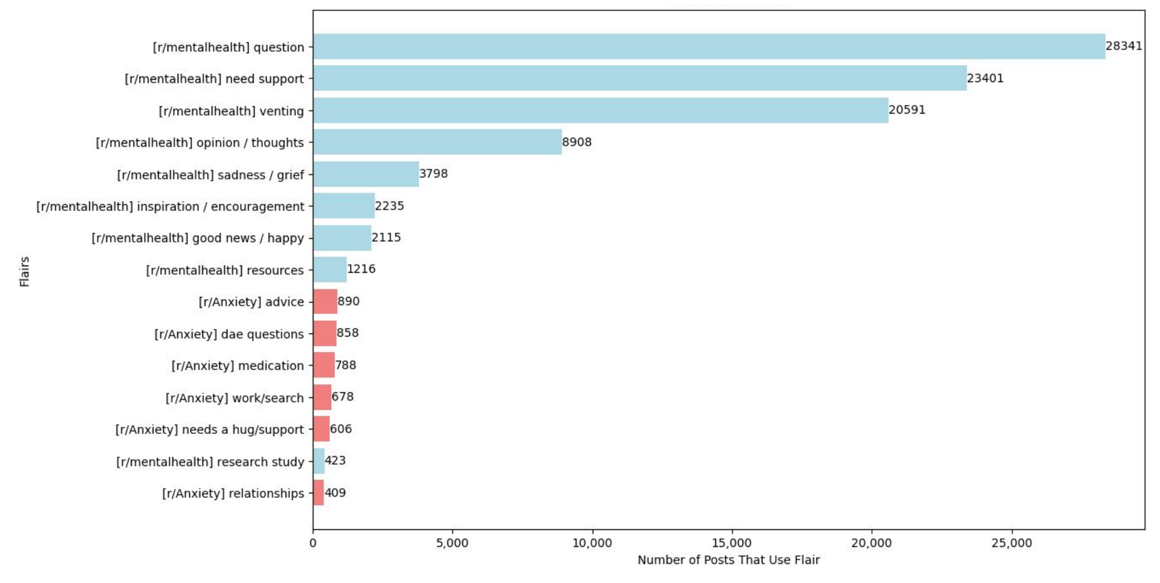
VADER – Sentiment Post Count Totals



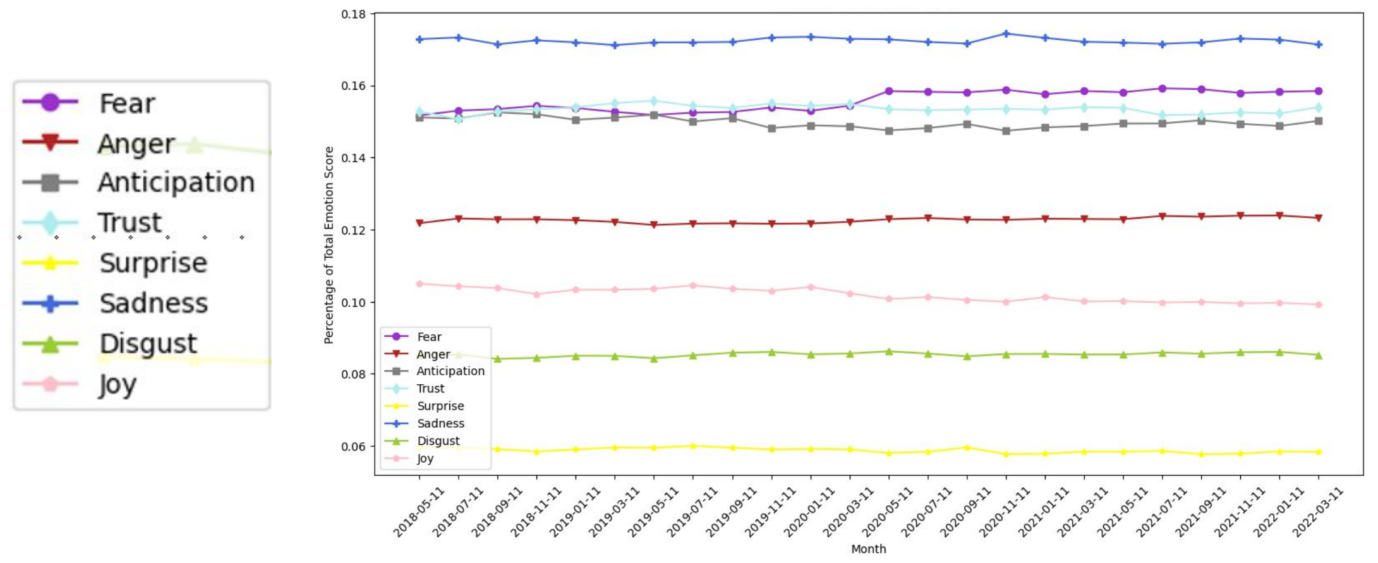
VADER – Sentiment Post Counts Over Time



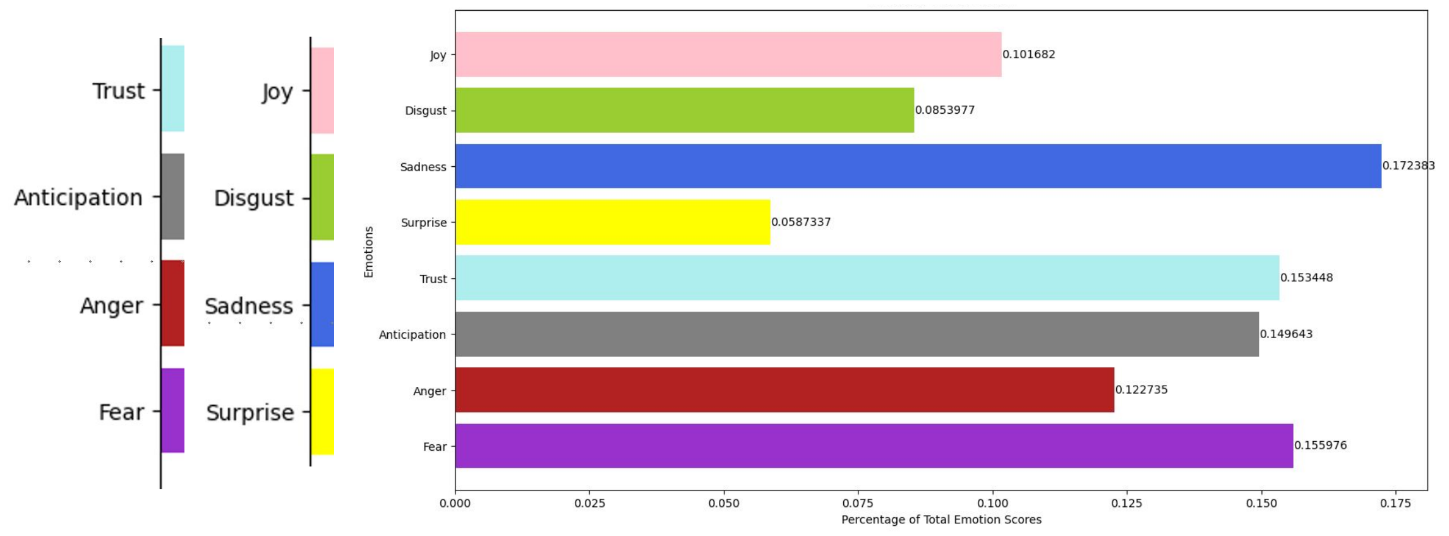
Flair Count Totals



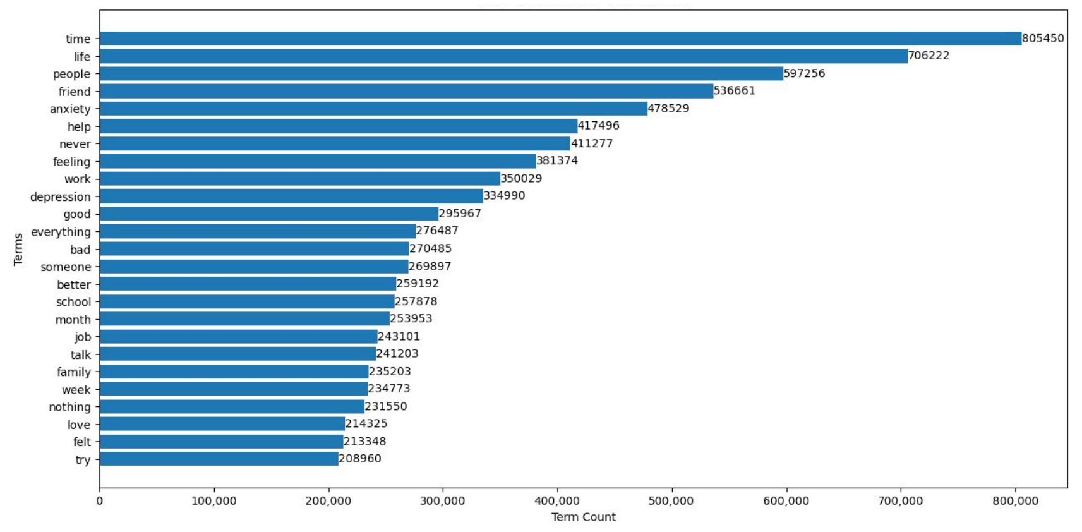
NRC Lexicon – Emotion Scores Over Time



NRC Lexicon – Emotion Score Totals



Term 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.