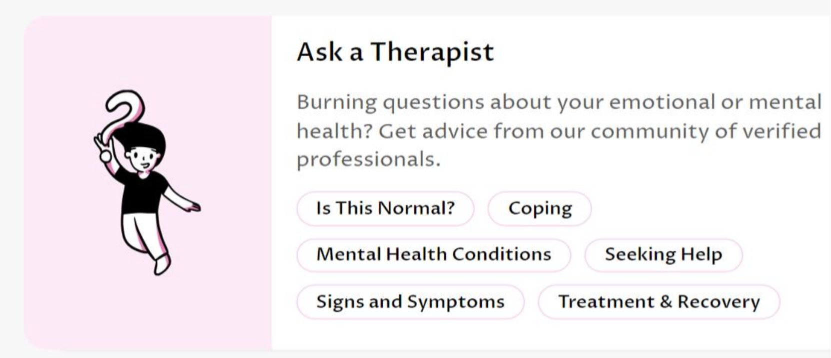
Automated Qualitative Thematic Analysis of Mental Health in Social Media Posts

Surya Appana, University of California, Berkeley (presenting author)
Ikhoon Eom, National University of Singapore
Jia Jun Tan, National University of Singapore

Ministry of Health, Office of Healthcare Transformation (MOHT), Singapore

Introduction

- Motivation:
  - Social media provides the opportunity for individuals to express mental health concerns.
  - Such concerns are of particular interest to healthcare organizations for making policy decisions.
- Singapore's **Ministry of Health, Office of Healthcare Transformation (MOHT)**:
  - Conducts broad-level healthcare analyses and projects to influence healthcare decisions in Singapore
  - Operates a social media platform, called **Let's Talk**, which is focused on mental health conversations & includes an interactive feature between individuals and therapists.



Ask a Therapist

Burning questions about your emotional or mental health? Get advice from our community of verified professionals.

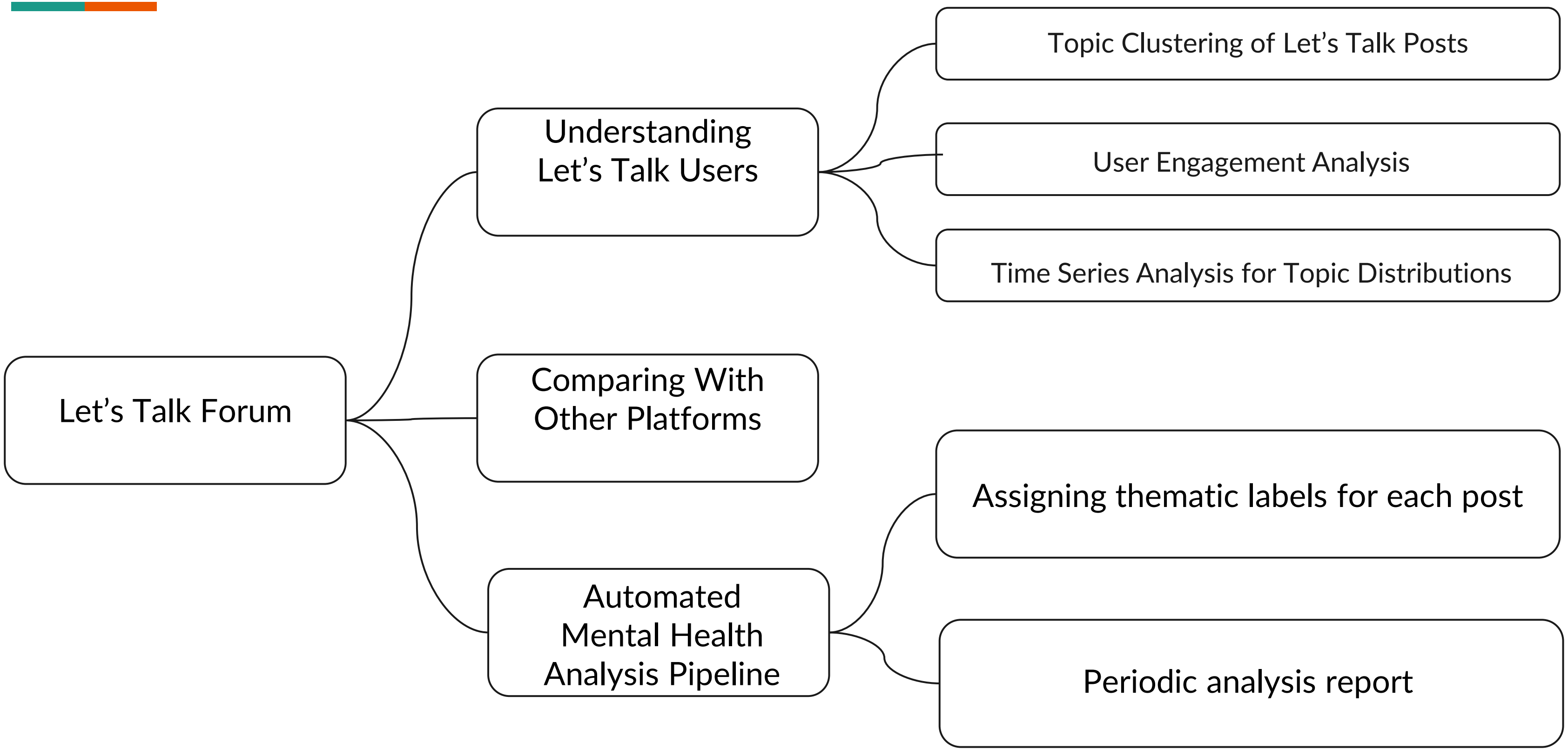
Introduction

- Unsupervised Natural Language Processing (NLP) through **topic modeling** provides new ways of discovering insights about mental health using social media text data.
    - Uncovering hidden topics and their distributions
    - Hidden structure related to mental health
- In this project, we aimed to discover such insights specifically as they relate to the goals of MOHT - mental health in Singapore

Objectives

1. Determine an optimal topic modeling algorithm for our dataset
2. Research questions Let's Talk posts
    a. What are the overall **topics** discussed on the forum, and which ones are most popular?
    b. Which topics **engage** users more?
    c. Do the topics **shift over time**? If so, are there trends?
3. Compare Let's Talk users with users from other forums:
    a. Are there any **differences between Let's Talk and other popular websites** in Singapore & other countries in terms of topics, user engagement, etc.?
4. Final goal: develop a **pipeline** by which practitioners can better understand the discussions and mental health of users on the Let's Talk forum and respond accordingly.

Project Overview



Let's Talk Forum

Understanding Let's Talk Users

Topic Clustering of Let's Talk Posts

User Engagement Analysis

Time Series Analysis for Topic Distributions

Comparing With Other Platforms

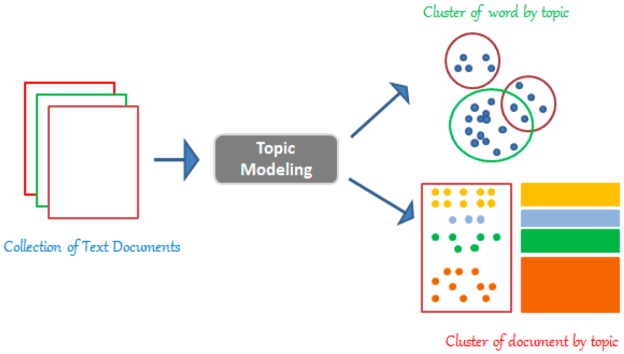
Automated Mental Health Analysis Pipeline

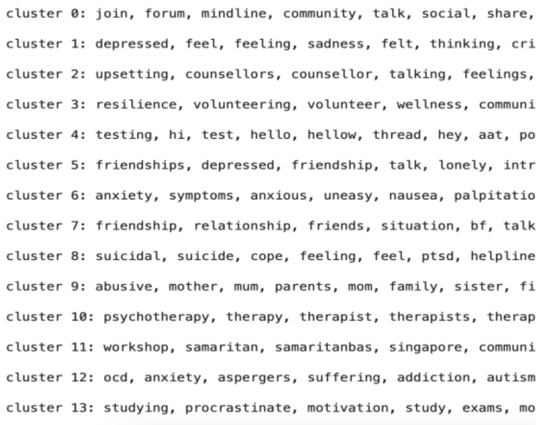
Assigning thematic labels for each post

Periodic analysis report

Methods

- Our task requires a **topic modeling** algorithm to cluster user posts together based on topic
- Four algorithms:
  - Latent Dirichlet Allocation (LDA): probabilistic model
  - Nonnegative Matrix Factorization (NMF): V = W * H
  - Top2Vec (embedding-based)
  - BERTopic (embedding-based)

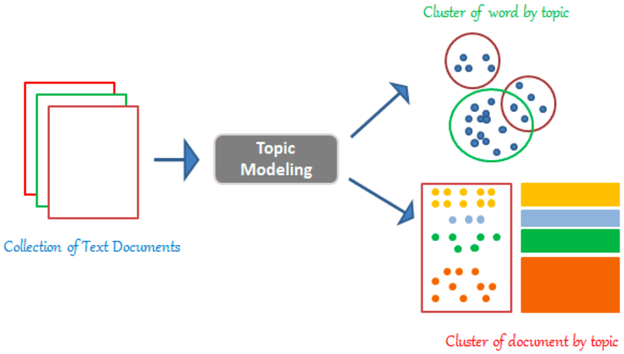


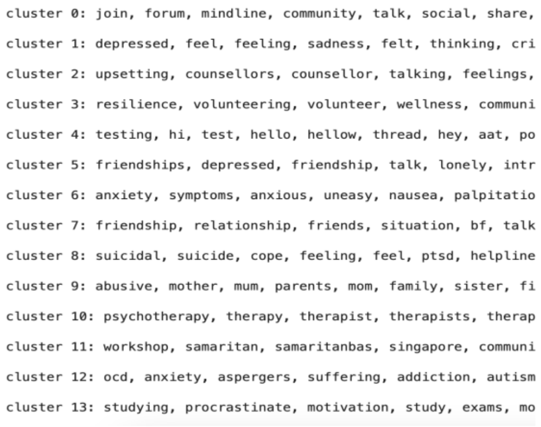


```
cluster 0: join, forum, mindline, community, talk, social, share,
cluster 1: depressed, feel, feeling, sadness, felt, thinking, cri
cluster 2: upsetting, counsellors, counsellor, talking, feelings,
cluster 3: resilience, volunteering, volunteer, wellness, communi
cluster 4: testing, hi, test, hello, hellow, thread, hey, aat, po
cluster 5: friendships, depressed, friendship, talk, lonely, intr
cluster 6: anxiety, symptoms, anxious, uneasy, nausea, palpitatio
cluster 7: friendship, relationship, friends, situation, bf, talk
cluster 8: suicidal, suicide, cope, feeling, feel, ptsd, helpline
cluster 9: abusive, mother, mum, parents, mom, family, sister, fi
cluster 10: psychotherapy, therapy, therapist, therapists, therap
cluster 11: workshop, samaritan, samaritanbas, singapore, communi
cluster 12: ocd, anxiety, aspergers, suffering, addiction, autism
cluster 13: studying, procrastinate, motivation, study, exams, mo
```

Methods

- Our task requires a **topic modeling** algorithm to cluster user posts together based on topic
- Four algorithms:
  - Latent Dirichlet Allocation (LDA): probabilistic model
  - Nonnegative Matrix Factorization (NMF): V = W * H
  - Top2Vec (embedding-based)
  - BERTopic (embedding-based)
- Final choice: BERTopic
  - Best DBCV score (Density-Based Clustering Validation)
  - Visual and qualitative analysis of topics yields the best results
  - Better able to capture semantic and contextual meanings through the BERT transformer

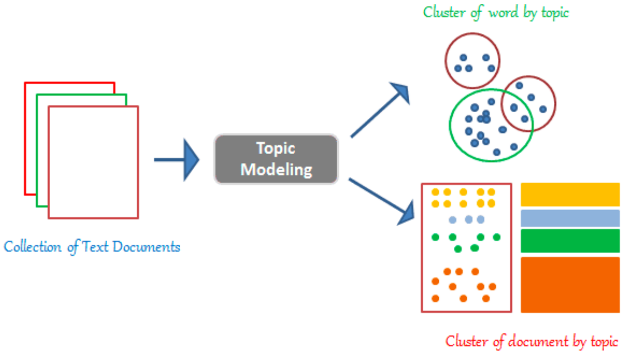




```
cluster 0: join, forum, mindline, community, talk, social, share,
cluster 1: depressed, feel, feeling, sadness, felt, thinking, cri
cluster 2: upsetting, counsellors, counsellor, talking, feelings,
cluster 3: resilience, volunteering, volunteer, wellness, communi
cluster 4: testing, hi, test, hello, hellow, thread, hey, aat, po
cluster 5: friendships, depressed, friendship, talk, lonely, intr
cluster 6: anxiety, symptoms, anxious, uneasy, nausea, palpitatio
cluster 7: friendship, relationship, friends, situation, bf, talk
cluster 8: suicidal, suicide, cope, feeling, feel, ptsd, helpline
cluster 9: abusive, mother, mum, parents, mom, family, sister, fi
cluster 10: psychotherapy, therapy, therapist, therapists, therap
cluster 11: workshop, samaritan, samaritanbas, singapore, communi
cluster 12: ocd, anxiety, aspergers, suffering, addiction, autism
cluster 13: studying, procrastinate, motivation, study, exams, mo
```

Methods

- **Bidirectional Encoder Representations from Transformers (BERT):**
  - A well known transformer baseline for NLP experiments
  - Able to capture semantics, emotions, and meanings behind various vocabularies within their respective contexts.
- BERTopic:
  - Embeds text into vectors using a pre-trained BERT model (trained on a massive Reddit dataset) & performs dimensionality reduction (nonlinear) on these vectors
  - Uses the embeddings to cluster documents (posts) into groups (using hierarchical clustering)
  - Outputs a topic representation for each group (sequence of words) indicative of each group's topic.

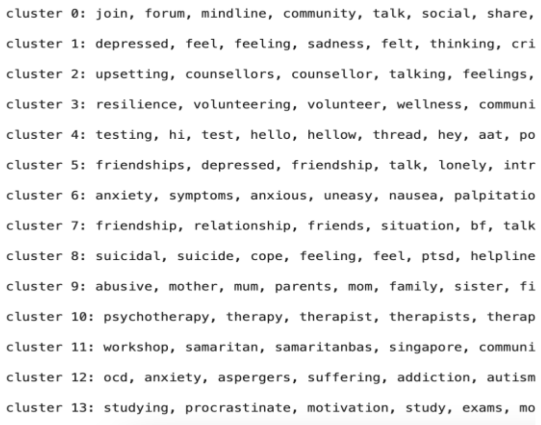


Cluster of word by topic

Collection of Text Documents

Topic Modeling

Cluster of document by topic

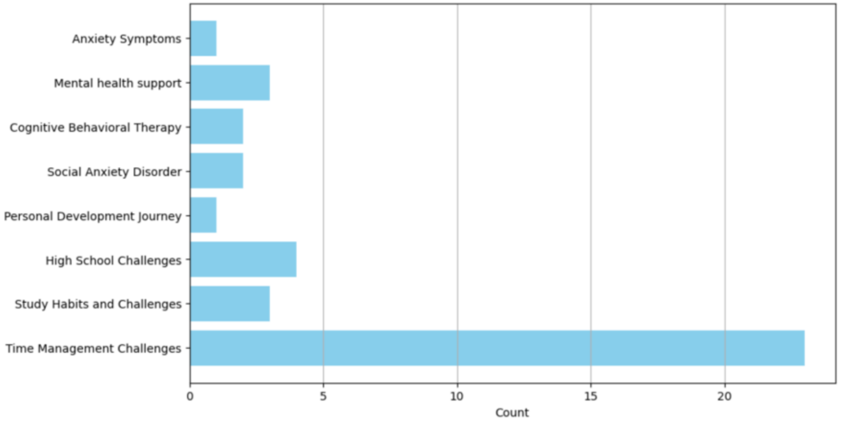


```
cluster 0: join, forum, mindline, community, talk, social, share,
cluster 1: depressed, feel, feeling, sadness, felt, thinking, cri
cluster 2: upsetting, counsellors, counsellor, talking, feelings,
cluster 3: resilience, volunteering, volunteer, wellness, communi
cluster 4: testing, hi, test, hello, hellow, thread, hey, aat, po
cluster 5: friendships, depressed, friendship, talk, lonely, intr
cluster 6: anxiety, symptoms, anxious, uneasy, nausea, palpitatio
cluster 7: friendship, relationship, friends, situation, bf, talk
cluster 8: suicidal, suicide, cope, feeling, feel, ptsd, helpline
cluster 9: abusive, mother, mum, parents, mom, family, sister, fi
cluster 10: psychotherapy, therapy, therapist, therapists, therap
cluster 11: workshop, samaritan, samaritanbas, singapore, communi
cluster 12: ocd, anxiety, aspergers, suffering, addiction, autism
cluster 13: studying, procrastinate, motivation, study, exams, mo
```

Methods

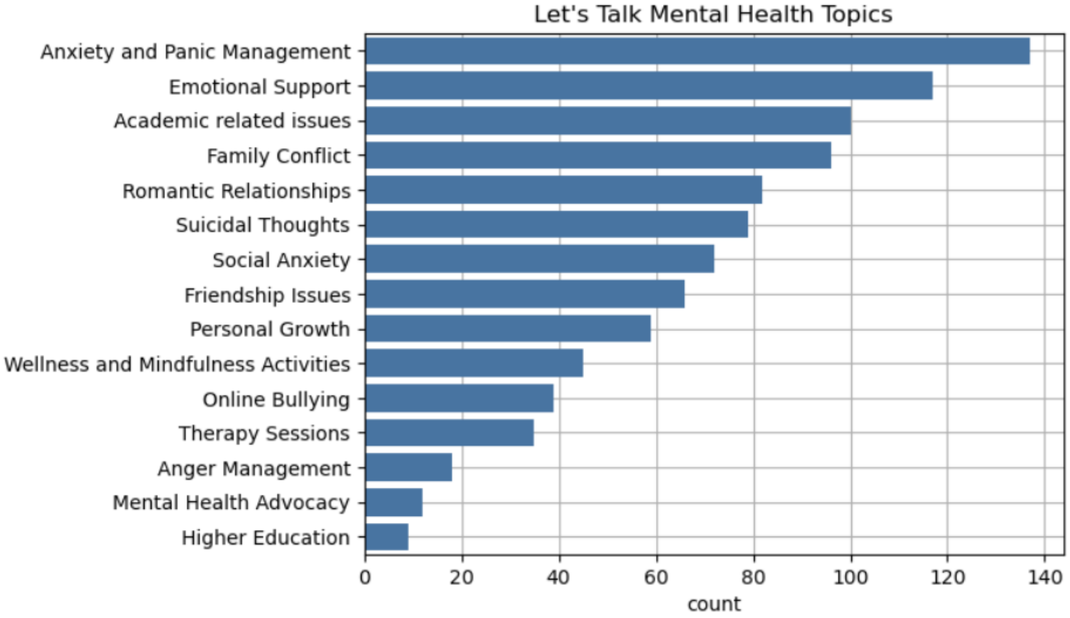
- BERTopic is stochastic:
  - UMAP (dimensionality reduction) is stochastic in nature
  - Resulting number and quality of clusters and topic representations are slightly different each time BERTopic is run
- **Question:** can we better distill the inherent topics after running our algorithm multiple times?
- Idea:
  - Aggregate all topic representations across many trials and cluster them based on a word frequency-based model - e.g. Nonnegative Matrix Factorization
  - Backtrack from these new clusters to determine the topic *distribution* of each post

Example Post and Topics



Topics Discovered on Let’s Talk Dataset

Summarizing the discovered topic representations into short descriptions, we came up with 15 distinct topics.

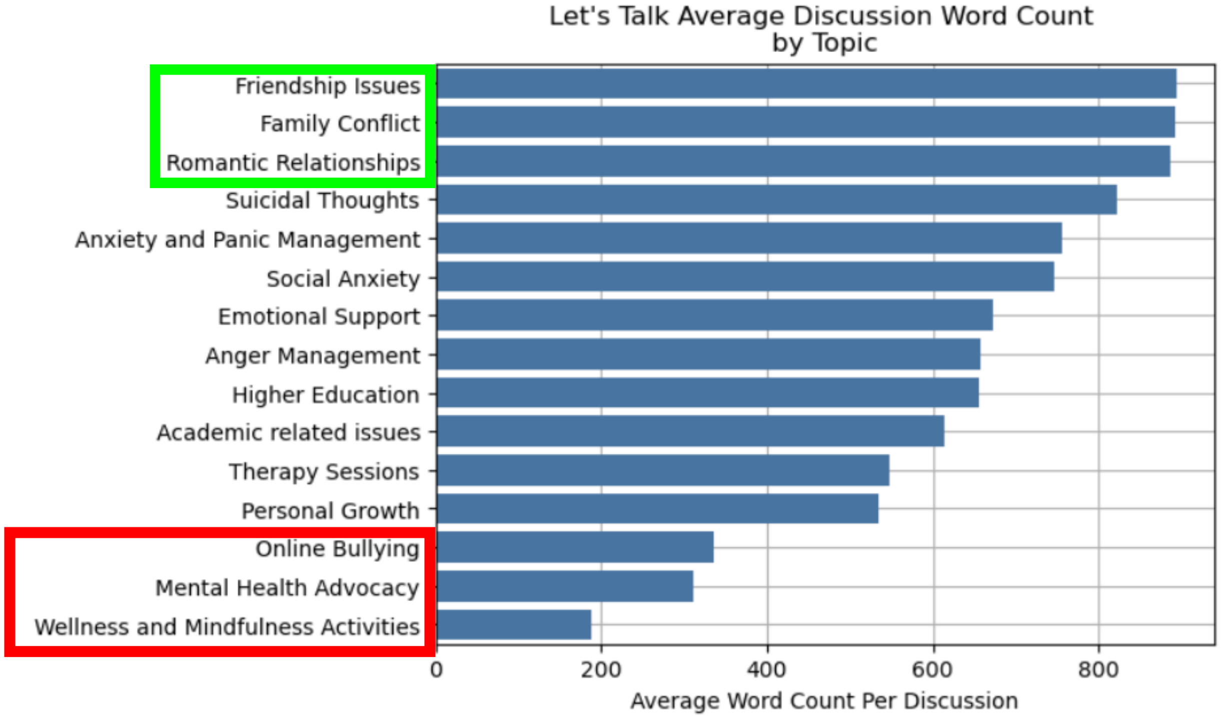


Examining User Engagement

- Having discovered the topics, we aim to answer:
  - 1) Are there differences in user engagement between individual topics?
  - 2) What are the topics individual users tend to focus on?
- We refer to engagement of users with the entire discussion following an initial post.
- Variables:
  - likes count, posts count, total word count, reads count, count of words in initial post
- Statistical test for differences between medians across topic groups:
  - Kruskal-Wallis test is significant at the 0.05 level for all of the above user engagement variables.
  - Total word count has the smallest p-value

Examining User Engagement

Q: Are there differences in user engagement between individual topics?

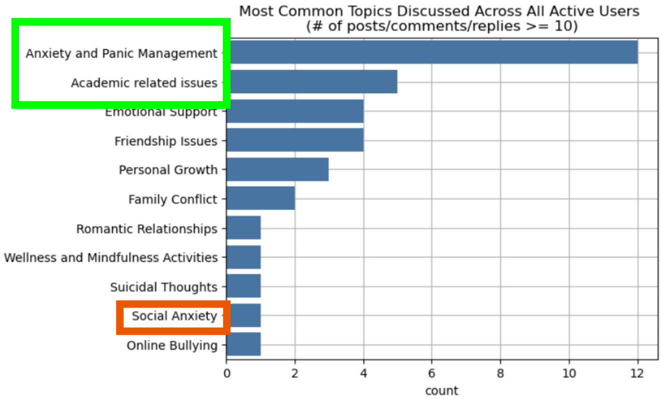


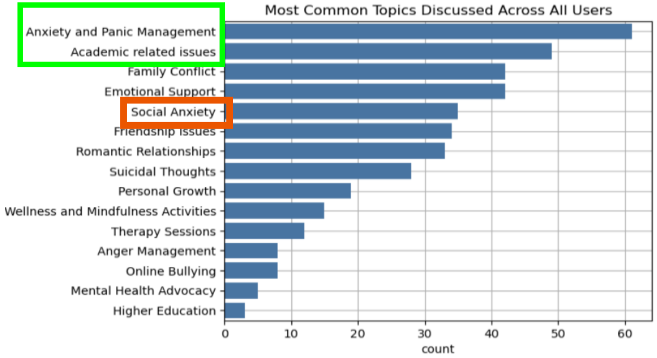
- Discussions about **friendship issues, family conflicts,** and **romantic relationships** seem to have the most total words on average

- Discussions about **online bullying, mental health advocacy,** and **wellness and mindfulness activities** seem to have the fewest total words on average

Examining User Engagement

Q: What topics do individual users tend to focus on? - For this, we care about the “active” users, which we define to mean users who have made a post, comment, or reply at at least 10 times on the entire





- Count refers to the number of users for which that topic is the topic they engage most in (across all of their posts/comments/replies)
- **Anxiety and Panic Management** and **Academic related issues** are the most common
- **Social Anxiety** drops in rank when we consider only active users
- The other topics approximately retain their ranks when considering only active

Comparison with Other Platforms

- Motivation:
    - Let’s Talk is relatively new
    - Investigating mental health in more popular platforms can inform future decisions on Let’s Talk based on new kinds of conversations that develop as Let’s Talk becomes more popular
- Platforms: (chosen based on data availability)
    - Reddit
    - HardwareZone -- Singaporean platform originally intended for computer tech related discussions, but has broadened over time

Datasets

- Reddit dataset:
  - ~80k posts after cleaning on Reddit mental health posts from a Hugging Face dataset that concern **adhd, depression, aspergers, ocd,** and **ptsd.**
  - We performed our topic clustering on a sample of 1500 posts (for each of adhd, depression, etc.) for convenience
- HardwareZone dataset:
  - Scraped ~500 posts using search keywords: e.g.
    - “anxiety,” “depression”
    - “mental trauma|struggling|stressed|difficult recover -News -Study -Survey”
    - “feeling stress life -News -Study -Survey”

Comparison with Other Platforms

- Several topics appear in the Reddit and HardwareZone datasets which do not exist as topics in the Let’s Talk data:
    - **Food/Nutrition** in Reddit data
    - **Drugs/substances** in Reddit data
    - **Job/Workplace stress** in HardwareZone
- There is only one **academic-related topic** that shows up (within the **ADHD** Reddit dataset), but for Let’s Talk, academic-related issues is the **third most common** topic.
    - ⇒ Academics is highly mentioned/discussed among Let’s Talk users, but this is not the case for Reddit and HardwareZone.

Possible Future Work

1. Soft clustering with Gaussian Mixture Models to determine topic probabilities for each post
2. Examining suicidal ideation for individual users to accelerate responses from therapists
3. Determining user intents for coming to Let’s Talk (or any other forum)

Thank You