AC 221 Final Project Report

Digital Moods: Investing the Connection between Twitter Activity and Mental Health

Siyuan (Tom) Zhang - Group 8

Introduction

The goal of my project is to investigate the relationship between Twitter usage and mental health. Twitter is used as the data source because it contains a vast amount of text that can be easily retrieved via its public Application Programming Interface (API). Mental health is the topic of interest here because it is a subject I care deeply about. There is plenty of research done regarding social media and mental health[1][2]. Here, I would like to conduct my own analysis, with an emphasis on Natural Language Processing (NLP) techniques and implementing an entire pipeline from scratch.

Methodology Overview

The pipeline is composed of several steps briefly described as follows.

1. **Data collection**: Write a data collection script that interacts with Twitter's API. Run it on Harvard's High Performance Compute Cluster (HPC) to scrape all tweets from a specific sample of Twitter users and time period.
2. **Sentiment analysis**: Use NLP techniques to analyze the sentiment of the tweets. Identify overall sentiment trends and variations between user groups.
3. **Topic modeling**: Build a topic model. Identify common themes and patterns in tweet content related to mental well-being.
4. **Domain-specific analysis**: Apply a filter to keep tweets with mental-health-related hashtags, then attempt to quantify mental health using a domain-specific technique.
5. **Statistical analysis**: Investigate intermediate results from the previous steps to uncover interesting patterns.

Data Collection

With a Harvard email, I was able to register for an academic research account on the Twitter Developer Portal which grants me access to their API. I am able to pull tweets from it with a monthly limit of 10 million tweets. However, it is worth noting that the API limits the number of tweets returned per request. Each response yields up to 500 tweets. Fortunately, one is able to retrieve the complete result set with pagination. In essence, each response contains a

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9915628/
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7364393/

“next_token” field which points to the next partition of the complete result set. Hence, by repeatedly sending API requests with an extra “next_token” parameter, we can pick up from where we left off. This is implemented with a loop in a Python script, which terminates when a particular response no longer has a “next_token” field.

Further, it is often good practice to handle any potential errors one may encounter when working with a web API. For instance, the data collection script could be exceeding the API’s rate limit of 300 requests per 15 minutes, in which case error 429 “too many requests” will be triggered. Similarly, 503 “service unavailable” can be encountered if the API is under maintenance. In these cases, the collection program will pause 30 minutes and attempt the last request once more.

As for query parameters, I specified the time period to be April and May, 2019. This is because May is Mental Health Awareness Month, and consequently I should expect to observe more relevant tweets during this month than any other. April will serve as a baseline which could be used for comparisons in analyses ahead.

What’s more, I had also questioned whether to collect tweets with mental-health keywords only. However, I realized that by focusing only on tweets with specific keywords or hashtags, I may miss out on other relevant content that does not include those terms. Additionally, this approach may lead to a smaller dataset, which might limit the generalizability of my findings. By contrast, collecting all tweets provides a more comprehensive view of the users’ activities during the selected time period. It allows me to analyze both mental health-related content and general content, which can help identify trends and correlations that might not be apparent when focusing only on mental-health keywords. Even though this would yield a significant amount of irrelevant content, we can mitigate this by simply collecting more data to ensure we still acquire a meaningful amount of mental-health-related content.

Finally, I considered whether to restrict the sample to tweets authored by celebrities only. The purpose is to further reduce the potentially large dataset. However, I did not pursue this path as it does not accomplish my research goal which is to draw conclusions on the general public. The mechanism to collect data should do its best to minimize bias and be representative of the target population.

Finally, as I have access to Harvard’s HPC through another class, I simply submitted a job to let the data collection script run remotely without the need to keep any local terminal sessions active. The script finished running in approximately 2 days and collected over 8 million tweets.

Sentiment Analysis

The next step of the pipeline is sentiment analysis. In essence, this is a process where we attempt to categorize the underlying sentiment given a piece of text. Traditionally, 3 levels of sentiments are used: positive, neutral, and negative. There are several models one could use to accomplish this classification task. Here, I have decided to experiment with the state-of-the-art NLP model: the transformer. It works by encoding a piece of text into word embeddings which are fixed-length numeric vectors that represent semantic meaning along with context. While it certainly is a powerful technique, the most obvious downside is that it is very computationally expensive to train a transformer from scratch, especially considering the size of our dataset. Fortunately, one can drastically reduce the training runtime with transfer learning.

Now, transfer learning is a useful technique that allows a researcher to borrow a pre-trained model (usually trained on very large datasets and are made publicly available), then we fine-tune the model with task-specific data relevant to our problem at hand.

To perform transfer learning, one requires labeled data with predetermined sentiment levels. Since classifying by hand requires too much effort, I searched online and found that a group of students at Stanford had already accomplished the task[3]. Conveniently, they were also working with Tweet data. With the labeled dataset, we are now able to commence transfer learning. I chose a general-purpose pre-trained transformer BERT (Bidirectional Encoder Representations from Transformers), which is publicly available on the model-hosting platform Hugging Face[4]. Unfortunately, the training process is nonetheless extremely slow: I waited 3 days, yet the model was still fitting on the first epoch (out of 10 epochs).

As a last resort, I searched for pre-trained transformers that had already been fine-tuned. Conveniently, the NLP group at Cardiff University had one of these models on Hugging Face[5]. This model is fine-tuned on tweet data specifically for sentiment analysis tasks. The model is based on the RoBERTa architecture, which is an improved version of BERT. Since it has been fine-tuned on tweet data, it is expected to perform well on my particular task. With the model ready, all that's left is to apply it on each piece of Tweet in my dataset. The process completed in just over a day.

Topic Modeling

Next I conducted topic modeling, a text analysis technique for extracting the abstract topics that occur in a collection of documents (in our case, Tweets). Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. For the purpose of

[3] http://help.sentiment140.com/for-students
[4] https://huggingface.co/distilbert-base-uncased
[5] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

my project, I will use it to uncover the salient points associated with mental-health-related Tweet content.

Latent Dirichlet Allocation (LDA) and BERT are two of the most popular approaches for this task. LDA is a generative probabilistic model specifically designed for topic modeling, while BERT is a pre-trained transformer capable of various NLP tasks. Here, I decided to compare both and implement a model with each technique. It is worth noting that the two models are quite different with various advantages and disadvantages.

Methodology-wise, LDA is the more traditional choice. It is an unsupervised machine learning model that generates topic distributions for documents and word distributions for topics. LDA works by assuming that each document in a corpus is a mixture of a fixed number of topics, and each topic is a mixture of words. In my implementation, I set the maximum topics to 15. On the other hand, BERT is the newer, state-of-the-art NLP model. To use BERT for topic modeling, I would need to fine-tune it on a labeled dataset and use clustering and dimensionality reduction on the output embeddings.

Performance-wise, LDA does not consider context whereas BERT is context-aware, so I expect BERT to produce more accurate outputs.

Nonetheless, the single major downside of BERT is that it is extremely computationally expensive. Note that on the HPC, I do not have access to GPU's; hence, any attempt to accelerate the training process is by parallelizing with CPU cores. The following are results from running a few experiments:

- 100 Tweets; 45 seconds
- 200 Tweets; 80 seconds
- 300 Tweets; 112 seconds

Assuming that the runtime increments linearly, it is estimated that it would take over a month for the model to finish training. LDA, by contrast, finished within 2 days, making it the more practical choice.

Domain-Specific Analysis

Something else I was interested in was to quantify mental health using domain-specific methods. One such example is the theory-driven measurement technique known as psychometrics[6]. In essence, a subject is given a questionnaire composed of questions with numeric rating scales; a psychologist can then evaluate their scores to gain valuable insights into the subject.

[6] https://en.wikipedia.org/wiki/Psychometrics

In the scope of NLP, one can potentially automate this process by encoding the psychometric scale items (questions) and a subject's response (a Tweet in our case) using a transformer. The two segments of text are now both converted to word embeddings. For each pair of Tweet and scale item, we calculate a cosine similarity score; we then average the scores. Now we have a numeric result that quantifies how close a Tweet is related to the given psychometric scales. This process is a novel text analysis technique known as Contextualized Construct Representation (CCR)[7].

I decided to use two questionnaires in my analysis: Beck's Depression Inventory (BDI) and General Anxiety Disorder-7 (GAD-7). As the names suggest, BDI measures depression and GAD-7 measures anxiety. The following figure is an example questionnaire used by actual psychologists.

GAD-7 Anxiety

Over the last two weeks, how often have you been bothered by the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Feeling nervous, anxious, or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Worrying too much about different things	0	1	2	3
4. Trouble relaxing	0	1	2	3
5. Being so restless that it is hard to sit still	0	1	2	3
6. Becoming easily annoyed or irritable	0	1	2	3
7. Feeling afraid, as if something awful might happen	0	1	2	3

Column totals _____ + _____ + _____ + _____ =

Total score _____

Figure 1: GAD-7, the psychometric scale questionnaire

Due to the domain-specific nature of this task, I applied CCR on a subset of Tweets with mental-health-related hashtags only:

[7] https://osf.io/bu6wg/

['#mentalhealth', '#mentalhealthawareness', '#mentalhealthmatters', '#selfcare', '#selflove', '#anxiety', '#depression', '#ptsd', '#bipolar', '#ocd', '#eatingdisorders']

This reduced the dataset to approximately 5500 Tweets.

Statistical Analysis

Now that I have acquired sentiment labels, topic labels, and psychometric cosine similarity scores, I can conduct a variety of statistical analysis procedures to investigate correlations and discover interesting patterns.

First, I visualized the distribution of sentiments across all Tweets in general as well as the mental-health-related hashtag-filtered Tweets.

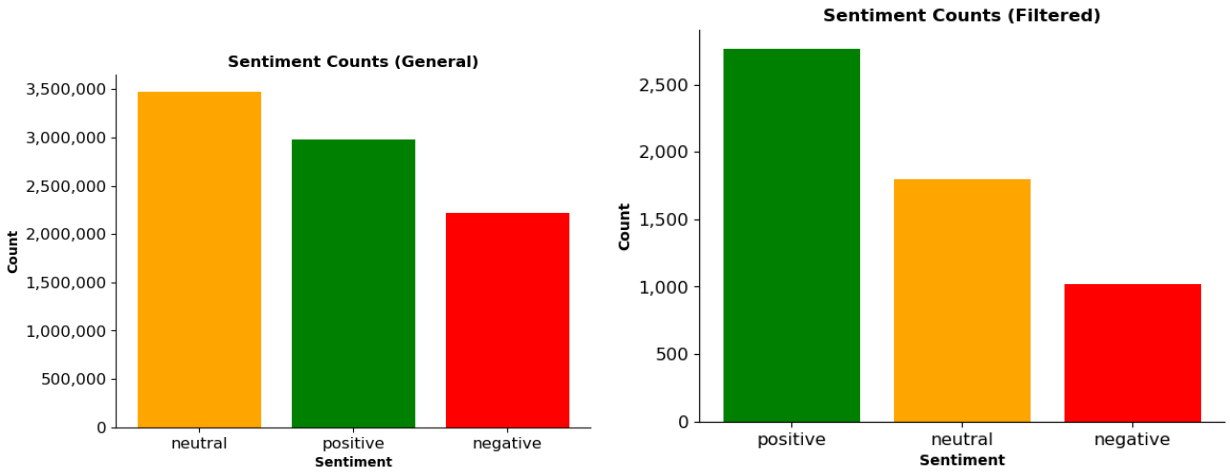

Figure 2: Sentiment counts (general vs filtered Tweets)

It is observed that in general there are more neutral than positive sentiment tweets. However, for mental-health-related tweets, there are more positives than neutrals.

I investigated further as I was curious whether there were any time-series trends of sentiment levels.

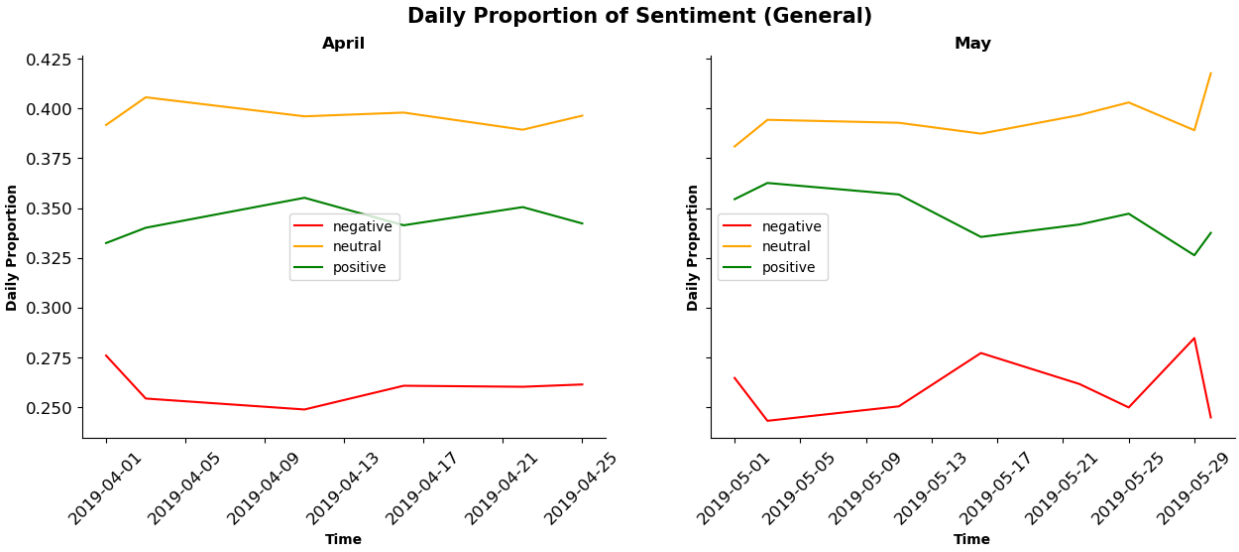

Figure 3: Daily proportion of sentiment (general)

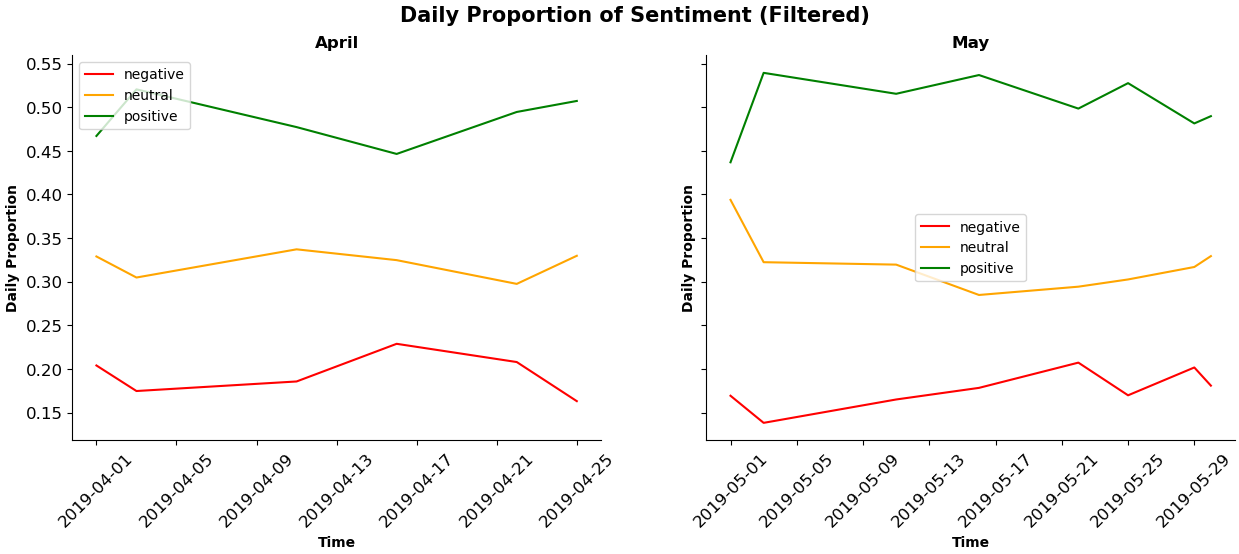

Figure 4: Daily proportion of sentiment (filtered)

I observed that April is quite stable sentiment-wise for both the general and filtered datasets. In May, we see a decline in the proportion of positive tweets and increase in negative tweets per day; this pattern is present in both the general and filtered Tweets.

Next, I plotted distributions of topic labels for the two groups, divided into April and May.

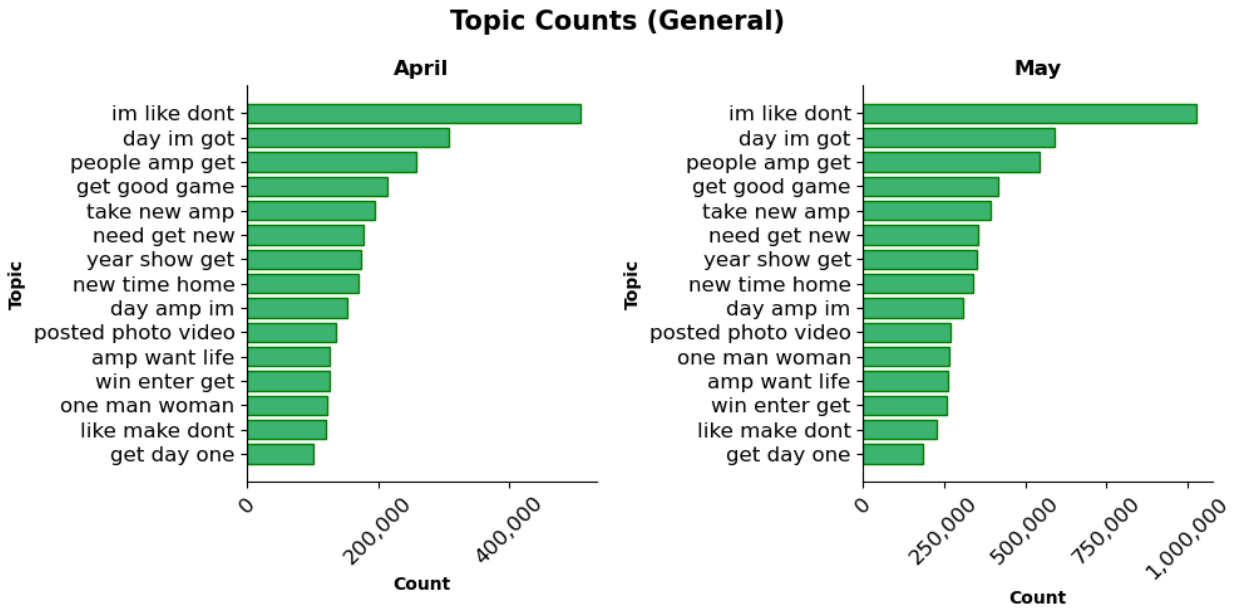

Figure 5: Topic counts by month (general)

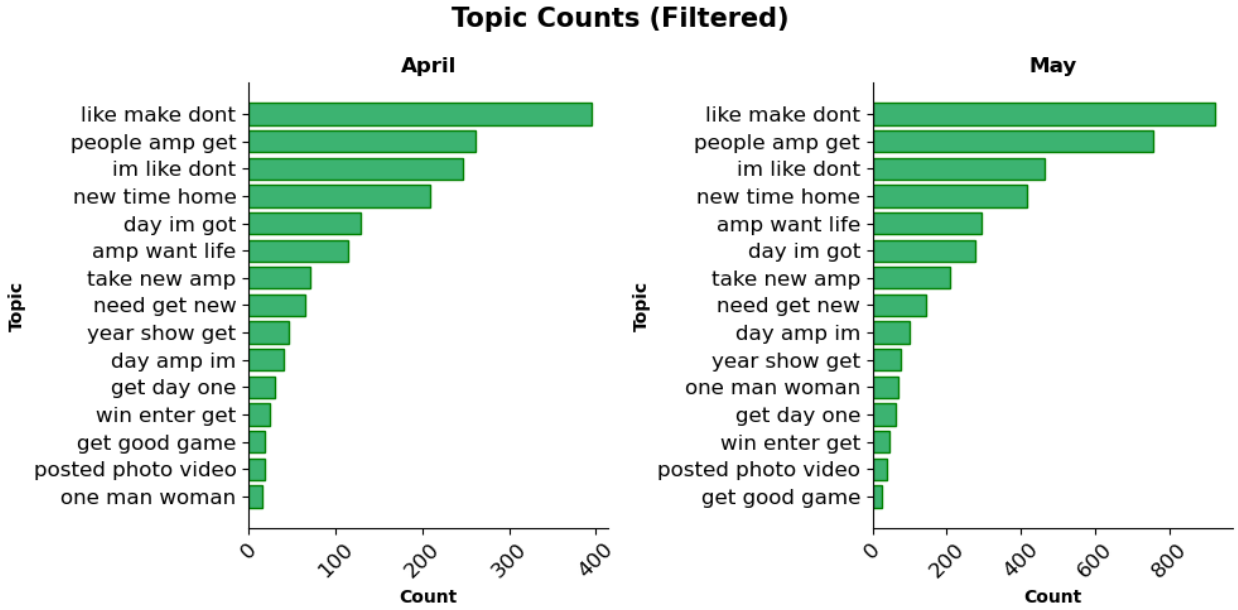

Figure 6: Topic counts by month (filtered)

We notice that the topics picked up by LDA are quite colloquial. And it is clear that most people often feel upset (e.g., the topic labeled “I’m like don’t”).

Similarly, we also investigate time-series trends here. Because visualizing all 15 topics in a single plot can be overwhelming, I kept the top 5 most popular topics overall. I then further calculated their proportions with respect to all available topics each day.

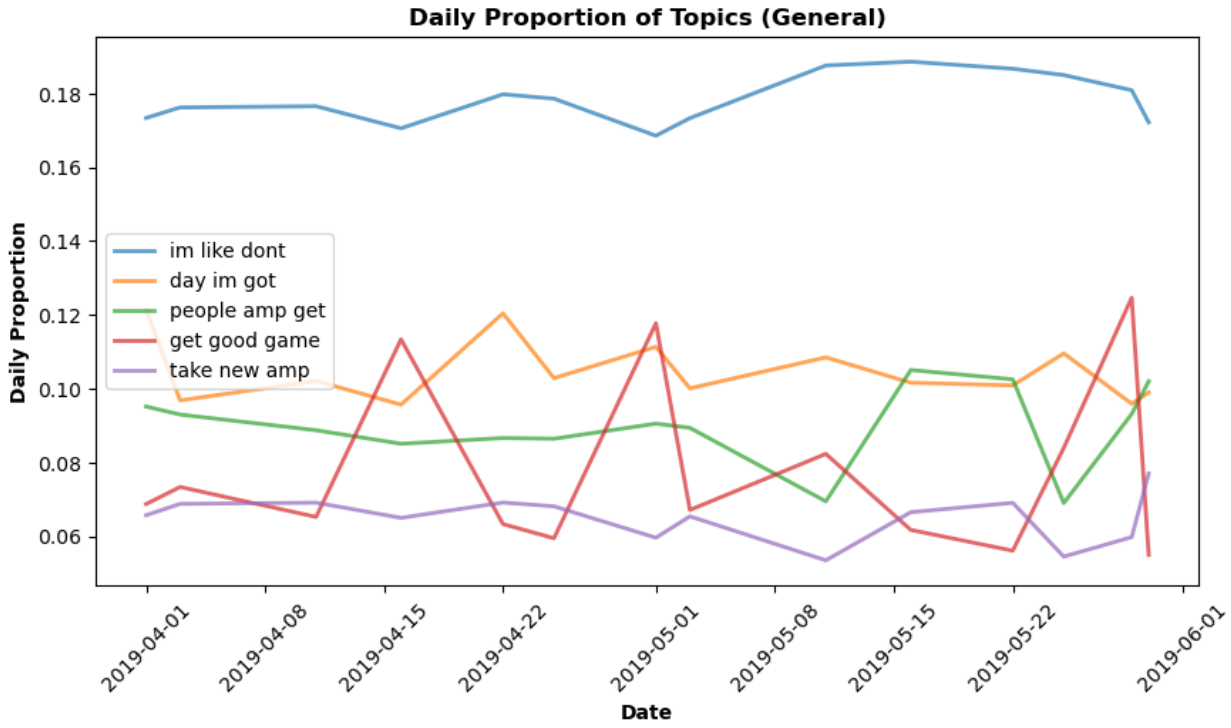

Figure 7: Daily proportion of topic (general)

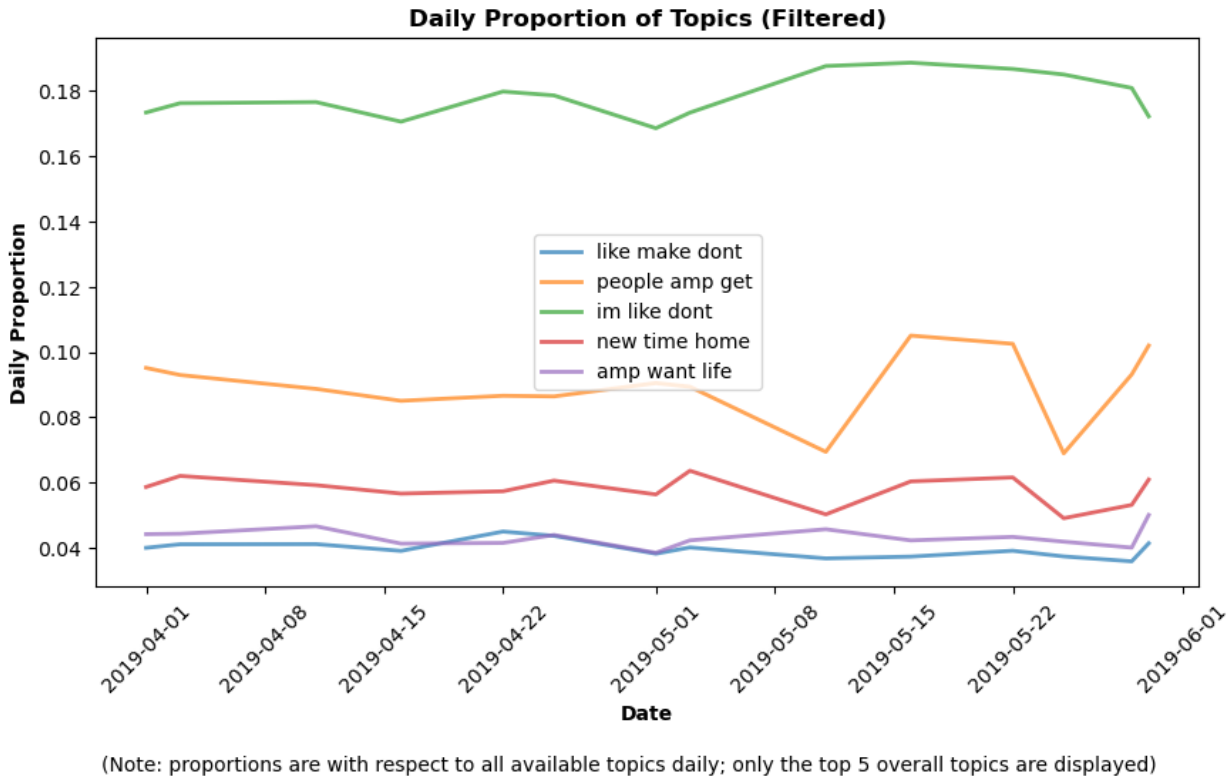

Figure 8: Daily proportion of topic (filtered)

The most popular topic ("I'm like don't") for both groups appears to be slightly increasing in proportion over time. Beyond that, there are no noticeable patterns.

Finally, we inspect our CCR results by calculating the correlation between sentiment and the psychometric cosine similarity averages for each questionnaire. Note that the sentiment labels are categorical, so we first convert them to integers (-1, 0, 1). For BDI, the correlation is -0.025765; for GAD-7, the correlation is 0.023042. Both are close to zero. This is likely due to the fact that the filtered dataset is not large enough, and it is possible to amplify the effect if more data were collected.

But if one were to interpret the correlations, keep in mind that nearly all the questionnaire items describe mental-health-related symptoms. A large cosine similarity between a Tweet and the questionnaire could indicate that the individual likely suffers from those conditions. Because we also converted sentiment levels to a sequential numeric scale, large values indicate positive sentiments. Hence, cosine similarity scores and sentiment values move in opposite directions. Looking back at the correlation between sentiment and average BDI cosine similarity–since the correlation is negative, we can conclude that negative sentiments correspond with depression symptoms, intuitively. However, the GAD-7 correlation is positive which is a bit

counter-intuitive. On the other hand, this observation testifies to the fact that an individual may suffer from anxiety even if their sentiment is positive.

Ethics Discussion

Here I will discuss the ethical implications I have considered for my project as well as acknowledge any potential ethical risks.

First, even though I collected geographical data and made quite an effort to handle the special case of geo-tagged Tweets during the collection process, I decided not to incorporate this piece of information into subsequent analysis out of privacy concerns. Over the semester, I learned that every piece of data could act as an indirect identifier no matter how trivial they seem. At the end, I retained only the bare minimum of text content and model outputs. Indeed, even the Tweet text can be used as identifiers too if the composition style is unique enough; however, I do require text to perform NLP procedures.

Throughout the project pipeline, I encountered several runtime-related challenges and were forced to adopt seemingly undesirable options such as using a pre-trained sentiment model and a context-unaware topic model. These tradeoffs were certainly not ideal, making the reliability of any subsequent findings questionable; nonetheless, compromises were still made due to practical reasons.

Further, mental-health diagnostics is a serious matter; there exist medical practitioners who are trained specifically for this task. Hence, the NLP technique used in the project (CCR) is not meant to replace clinical diagnosis. It is only an assistive screening method to facilitate the process at best.

Code

The code written for this project can be accessed in my public Github repository: https://github.com/tomzhang255/AC221_Project