# Assignment 2

Saujanya Acharya Eric Ohemeng Bozhou(Peter) Tan

2024-03-24

## Introduction

In this project, the research topic we have chosen is whether the current economic situation in the United States is improving or deteriorating. As we enter 2024, some surveys and statistical data indicate that the U.S. economy is moving in a positive direction. For instance, the Consumer Sentiment Index has risen by over 20 points since last December, and inflation expectations have also decreased, falling into the range seen in the past few years. However, many individuals also express that they haven't felt the improvement in the economic situation themselves; instead, they feel it's getting worse — with inflation still high and income growth sluggish. We aim to understand the perspectives of different individuals in society regarding the current economic situation through discussions among users on Reddit.

## Methods

To extract the threads and comments from Reddit, we first need to choose the keywords or subreddits. There are two methods to extract the posts through Reddit API: one approach is searching the keywords to get a list of threads and extracting comments from those threads; the other way is choosing a specific subreddit and extracting the comments from the subreddit. The former method provides the threads from a wide range of subreddits, which means there are lots of irrelevant threads in the results; the latter method can make sure all threads are under a specific topic or keyword. After several experiments, we adopt the latter method, choosing several subreddits for listening.

Economy situation is a large topic, including inflation rate, salary and unemployment rate. First of all, we choose the subreddit "economy", because this subreddit contains threads and comments covering almost all topics about economy, which can help us understand people's feelings about economy situation from many perspectives. There are also hundreds of comments in this subreddit, making the sample size large enough. Secondly, we also choose some subreddits focusing on a specific economic topic: "inflation", "business" and "salary". These subreddits are also helpful but there are some disadvantages compared to the subreddit "economy". Firstly, threads from these subreddits mainly focus on one economic topic. Secondly, lots of threads and comments are not related to our research topic. Therefore, while we also listen to these subreddits, we lean towards

utilizing comments from “economy”. If the data from “economy” is sufficient, we do not intend to utilize comments from other sources. If not, we will also reach records from them.

We listened to those four subreddits for 8 days, from March 16th to March 23rd. Every evening, one member in the group extracted comments from the subreddits posted in the last 24 hours, stored data as .csv files and uploaded them on **GitHub**. The details of listening results are shown in Table 1. There are at least 400 posts in “economy” subreddit every day and 4520 comments in total.

After collection, we divided the comments we have in “economy” subreddit into three parts and each member coded the comments into four categories: favor, oppose, neutral and irrelevant. Since the number of relevant posts from “economy” subreddit meets the requirement, we did not code the posts from other subreddits.

Table 1: The number of comments from different subreddits in 8 days

| Date | Economy | Inflation | Business | Salary |
|---|---|---|---|---|
| 2024-03-16 | 723 | 640 | 49 | 297 |
| 2024-03-17 | 513 | 185 | 293 | 73 |
| 2024-03-18 | 438 | 311 | 162 | 87 |
| 2024-03-19 | 551 | 339 | 94 | 128 |
| 2024-03-20 | 515 | 278 | 180 | 0 |
| 2024-03-21 | 795 | 1233 | 85 | 0 |
| 2024-03-22 | 571 | 328 | 116 | 0 |
| 2024-03-23 | 414 | 338 | 75 | 161 |

# Analysis

Building upon the comprehensive data collection framework outlined in the Methods section, we started the analysis of the posts by first cleaning the hence collected data. The posts with “Irrelevant” posts were first removed. Finally, we ended up with a list of 838 posts/comments that had some relevance to the topic of our interest.

The initial review of the hand-coded labels for each post showed that the majority expressed opposing sentiment to “economy is improving”. More than half of the posts indicated that the reddit users felt that the economy was not faring well, and has been worse than the past times. On the contrary, analysis indicated that less than 20% of the posts had positive/favorable sentiments associated to them.

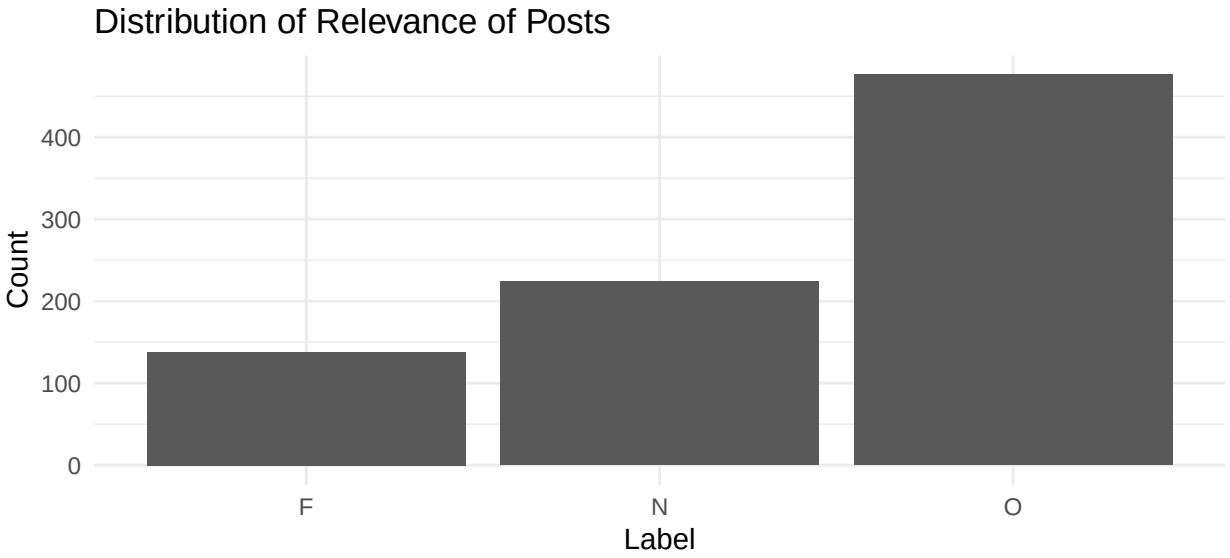


Figure 1: Distribution of in Favor (F), Neutral (N) and Opposing (O) sentiment of cleaned posts

Furthermore, the distribution of time of the posts over the day was also analyzed. The analysis illustrated how most of the posts were made on the 2nd half of the day through to near mid-night. The discussions were less frequent over the span of the night. Moreover, it was observed that the most frequent posts were made from 3 PM to 6 PM.

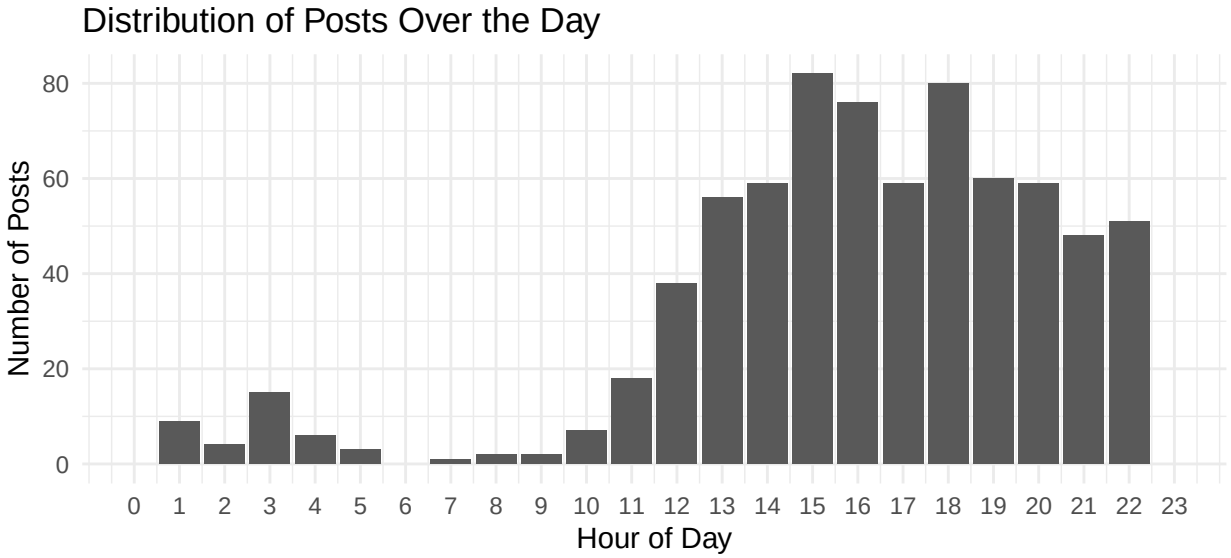


Figure 2: Distribution of posts made over the time of the day

After the initial meta-analysis, the posts were then reviewed to analyze the frequency of words to gather insights on specific topics that were being discussed in these forums. To do this, first the data was tokenized, and cleaned. Moreover, the common stop words were also removed with the use of tidytext library in R. It was observed that the most common words that appeared in the discussions were as expected from our initial assumption of the kind of discussions happening in the online forums.

The frequency of the most common/frequent words occuring in the comments has been visualized

below:

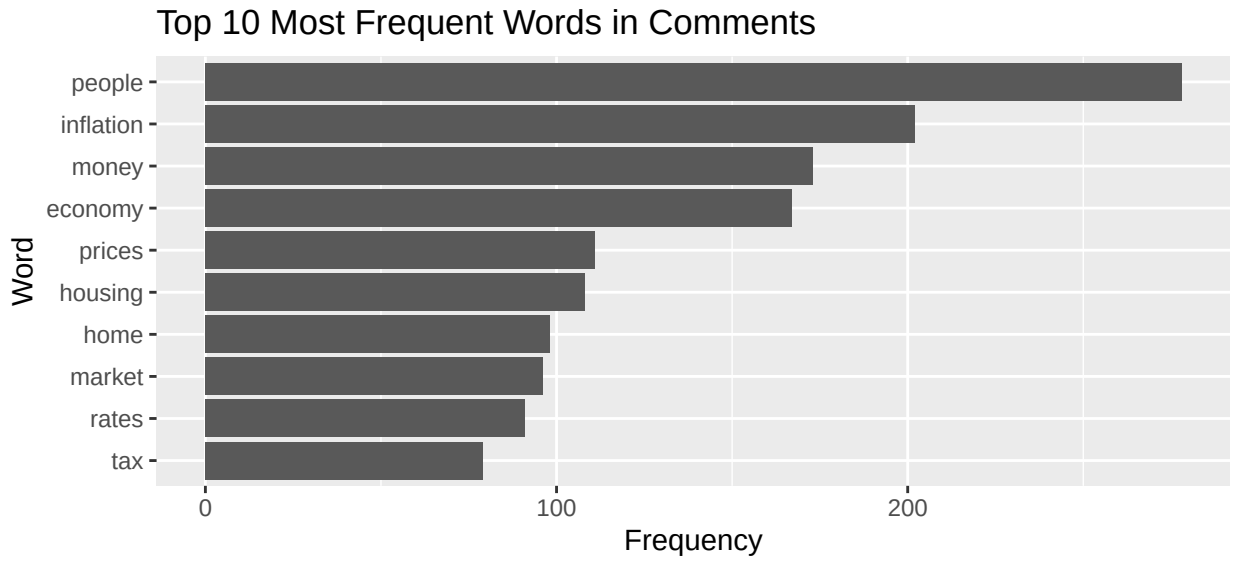


Figure 3: Frequency of most commonly used words in the discussions