Analyzing Social Trends on Recession conditions

Kanishk Barhanpurkar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
kbarhan1@binghamton.edu

Harshad Bhandwaldar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
hbhandw1@binghamton.edu

Nikita Mandlik
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
nmandli1@binghamton.edu

Brinda Eshwar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
beshwar1@binghamton.edu

ABSTRACT

Social Media platforms are widely used for transmitting data in different formats. We can generate a lot of information on the recent trends. Additionally, public opinions can be used as a tool to predict forthcoming events. In the post-covid era, we are facing many economic challenges like the recession and other economic crises. Social media monitoring tools use to give their users insights about how the public feels in regard to their business, products, or topics of interest and can help you see how positively or negatively your brand is perceived on social media based on the tone of mentions. We are collecting data from two social networking sites, Twitter and Reddit.We are also using News Articles as another data source.Twitter is a public social networking domain and Reddit is a community-based social media platform. The primary aim of this project is to analyze the social media-associated data which predicts the conditions which lead to the Recession over time. Secondly, for the third dataset, we will evaluate the news article which contains the topic of recession and economic crisis.

1. INTRODUCTION

Recession is a business cycle contraction when there is a general decline in economic activity. In 2020, Covid-19 spreads exponentially in the entire world due to which many nations imposed a lockdown on the entire nation. This decision affects international trade and decreases the currency flow among various countries [1]. Several countries whose economy is dependent on the tourism sector were affected badly. Therefore, the major output of this can be seen in 2022 which results in inflation and a lack of centralized money to fulfill citizens' basic needs. Social-networking sites like Twitter and Reddit play very vital roles in understanding public opinions [2]. The number of users using social networking sites makes it mandatory to analyze the trend of current affairs happening across the globe.

2. RESEARCH OBJECTIVE

In this project, we are using the Twitter Data (Twitter API), Reddit Data (Reddit API) as public-opinion data and NYTimes data as news article dataset for recession conditions. We are fetching the ID, text and timestamp from the twitter while we are collecting id, current datetime and postdate from analysis. In the New York Times API, we are using the abstract summary, headline, publication timeline (date, time and year) and word-count. We have pre-defined the topic of Recession and added few key-words as filter to the data scraping scriptd. Additionally, we will be filtering the tweets and reddit posts data based on the semantics of english languages. In this project phase, the main objective will be to collect the optimum amount of data, perform data cleaning and preprocessing and perform data analysis. We will be using different data visualization techniques for understanding the data better.

3. RESEARCH METHODOLOGY

We have divided this project phase in three steps based on the project description 2. These steps will be important to explore the research objectives in a correct manner. The steps are comprises of Data collection, Data cleaning and Result analysis.

3.1 DATA COLLECTION

We have already implemented the python scripts for the data collection of 3 datasets in Project phase 1 where we have used NYTimes API, Twitter API (continuouslyu streaming) and Reddit API for the collection of news articles, tweets and reddit data respectively [3] [4] [5]. We have used MySql for the data collection and all the scripts were implemented in Python language. For collecting twitter data, we have used Twitter stream API and for collecting Reddit data, we have used Reddit API. No high-level library is used such as PRAW and scrapy. We are also collecting data on Reddit as it is mention to provide information for the sub-reddit r/Politics in the Project Description 2.

3.2 DATA CLEANING

The collected data (tweet or reddit posts) contains many issues which are need to be addressed. When we first retrieved the Twitter data, it contains lot of URLs which are referring to other tweets. Additionally, removal of emoticons and stop words is also necessary because it will hinder the analysis results. In reddit data, we found out that the short forms used for texts like onl9 used for word online, JW used for just wondering and just wondering and ISO used for In Search Of needs to be addressed. Secondly, there are many mispelled data has been observed in the reddit data. For example, izz is used for “is”. The NYTimes data does not show these type of issues as the data is used as the source of information.

3.3 DATA ANALYSIS

We are using different data visulaization techniques for the better understanding of the data collected from various datasets. We will be using Matplotlib and Seaborn libraries for the graph plots. Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy [6]. Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data [7]. Also, we are using Wordcloud for the preliminary text analysis for every dataset. A wordcloud is a visual representation of information or data. It shows the popularity of words or phrases by making the most frequently used words appear larger or bolder compared with the other words around them [8].

4. DATA REQUIREMENTS

We will be collecting the 50,000 comments per week from the Reddit API and we will be getting 17,000 tweets per week for the keyword Recession. For NYTimes API, we are collecting 1800 new articles associated with economic crisis and recession conditions. We will continue collecting these data every hour for given timeline.

5. CONCLUSION

The major part of this project phase involve around data cleaning and data analysis part. We are prilmarily using data visualization techniques to understand the data. Additionally, we are identifying and rectifying the major issues associated with the data and filtering the impurities. In the future implementation, we will planned to perform the textual analysis techniques such as sentimental analysis.

REFERENCES

[1] Feyisa, H. L. (2020). The World Economy at COVID-19 quarantine: contemporary review. International journal of economics, finance and management sciences, 8(2), 63-74

[2] Mariolis, T., Rodousakis, N., & Soklis, G. (2021). The COVID-19 multiplier effects of tourism on the Greek economy. Tourism economics, 27(8), 1848-1855.

[3] TwitterAPI Documentation.
https://developer.twitter.com/en/docs/twitter-api

[4] RedditAPI Documentation.
https://www.reddit.com/dev/api/

[5] NYTimes API Documentation.
https://developer.nytimes.com/docs/articlesearch-product/1/overview

[6] Matplotlib Library Documentation
https://matplotlib.org/

[7] Seaborm Library Documentation.
https://seaborn.pydata.org/

[8] WordCloud Library Documentation.
https://pypi.org/project/wordcloud/