Influence of Economic Status on Social Media in Each State of United States

Team member:
Zhehu Yuan, Qifeng Zeng
Mucheng Luo, Fan Xu

Abstract

In this project, we try to find the influence of economic status on social media like Twitter, Instagram, in each state of the United States. We collect the posts data form several social media platforms such as Twitter, Instagram, and some check-in platforms over a period, we then combine these posts data with the economic data of each state over that time period. For each social media, we analyzed the relationship between the number of posts in each state in a month and the state's economic situation for that month.

1 Introduction

At present, there are many such official statistical data indicators to reflect economic situation. The most widely used indicator to measure economic performance is gross domestic product (GDP)[1], a term everyone is familiar with, which measures the market value of all final goods and services produced within a country over some period of time. There are also lots of smaller economic indicators, such as personal income, which refers to all the earnings made by a household in a given time. It includes various sources of income such as salaries, wages, investment, dividends, rent, contributions being made by an employer towards any pension plan, etc. But the acquisition of such traditional statistical data has its shortcomings, cause the statistical process is labor-intensive, time-consuming, complex and expensive [2]. Even the statistic of personal income such a small economic indicator includes so many aspects, not to mention the more extensive and more complex content that needs to be involved in the calculation of GDP. For example, the expenditure-based accounting sums up the purchases of goods and services by different groups or categories: consumption, investment, government expenditures, exports and imports.

So, we wanted to find a new indicator to measure the state of the economy. This kind of indicator is more intuitive in the current digital age, the statistics of data are easier and faster, and it can reflect the economic situation of a certain region at a certain point in real time, breaking through time lag and geopolitical constraints. In this project, we think posts from social media would be a good new indicator to reflect the economic situation. In order to verify our proposal, we need to find evidence to prove that there is a certain correlation between posts and economic situation.

We were inspired by a paper by Agustín Indaco[3], estimated economic activity by social media information. However, this paper only estimates GDP in U.S. with twitter information. Also, the analysis is done in 2012 which is obsoleted. So, we would like to estimate the effect with information on more social medias. Also, we want to see if we could still estimate economic activity by tweets information during COVID-19.

The goodness of our analysis is that we have different datasets from different social medias with some non-overlapping time period. To resolve the social media difference, we analyze data from different social media and compare them to see whether our conclusion have consistency. And we use the inflation rate to calculate real income [4], which better reflects the purchasing power of individuals regardless of the time factor.

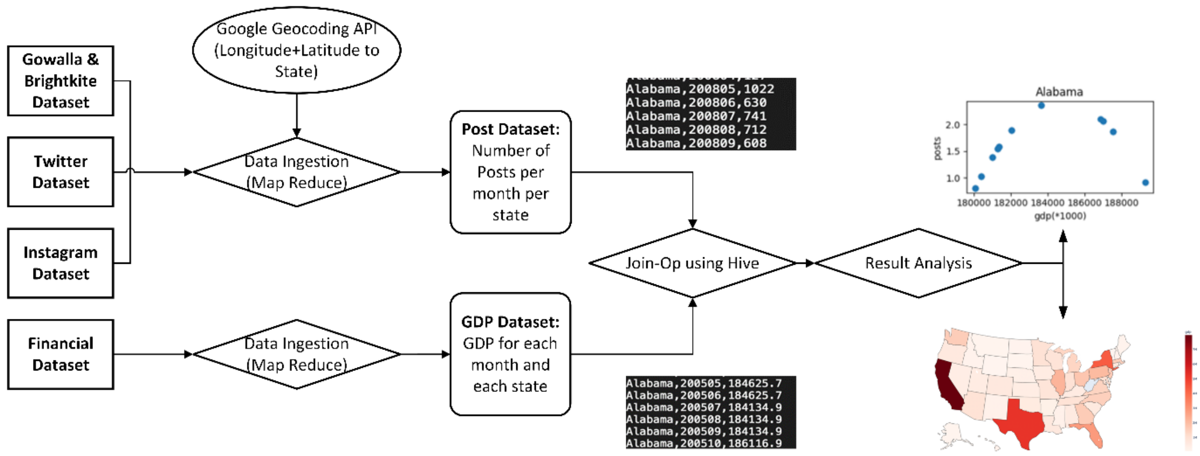

Figure 1 Design Diagram

2 Our Approach

The procedure we proposed for the project uses the post datasets from several social medias which contains posts with the date and locations and the financial dataset that contains the economic status data for each state in the United States over a long period of time to find the potential influence of the economic status on social media. We'll discuss the detailed dataset in the next section.

As shown in Figure 1, we first collected the dataset over the Internet or from organizations and saved them in the Hadoop Peel Cluster. We then used Map-Reduce to do the data ingestion where we cleaned the data, excluded the data with missing attributes, and extracted the posts happened in United States. After the ingestion, we got the posts dataset which is of the format:

State, YYYYMM, number of posts

and the GDP dataset which is of the format:

State, YYYYMM, GDP of that month

Using Hive via the HPC, we did a Join operation that combined both dataset if the state and time are the same. With the combined dataset, we did an OLS regression to analyze the influence of economic status. We present a sample of the final dataset and output in Figure 1.

It's worth mentioning that the datasets we collected are from different social medias with some non-overlapping time periods. To resolve these differences, we analyze data from different social medias separately, and use the inflation rate to calculate real income, which better reflects the purchasing power of individuals regardless of the time factor. We plan to get results for each social media and compare them to see whether our conclusion have consistency.

In the following section, we will discuss the dataset, ingestion, and experiment setup in detail.

3 Experiment

3.1 Dataset

We have collected four datasets, and we'll introduce them in detail with a sample of each dataset shown in Figure 2.

3.1.1 Twitter

The twitter dataset we used is called GeoCoV19, a large-scale Twitter dataset related to the ongoing COVID-19 pandemic. The dataset has been collected over a period of 90 days from February 1 to May 1, 2020 and consists of more than 524 million multilingual tweets with the users' geolocations. The dataset is about a few hundred of GB in size and the data is in JSON format.

3.1.2 Instagram

We collected the Instagram dataset from X-Byte. Instagram is one of the most famous social media. This dataset contains Posts data with detailed attribute including date, time, state, number of likes, and so on. The users are from 31 states in the United States. The dataset is about 5 GB in size.

3.1.3 Check-in dataset

The check-in datasets are first introduced by Stanford University. They collected the check-in data from two location-based social networking service providers, Gowalla and Brightkite.

```
{"tweet_id":"1223489811459108864",
"created_at":"Sat Feb 01 06:14:37 +0000 2020",
"user_id":"29512878",
"geo_source":"user_location",
"user_location":{"country_code":"ph","state":"Cavite","city":"Dasmarinas"},
"geo":{},
"place":{},
"tweet_locations":[]}
```


Twitter dataset sample


```
{"review_id":"fj7N9Lp6AvEEy6LHrDZzjw",
 "user_id":"gy7Ss1uTpCjbbGsghTvNsw",
 "text":"Midwest Cannabis Seeds è un... #cannabislife",
 "date":"2019-07-29 17:12:27",
 "state":"PA"}
```


Instagram dataset sample

[user]	[check-in time]	[latitude]	[longitude]	[location id]
196514	2010-07-24T13:45:06Z	53.3648119	-2.2723465833	145064
196514	2010-07-24T13:44:58Z	53.360511233	-2.276369017	1275991
196514	2010-07-24T13:44:46Z	53.3653895945	-2.2754087046	376497
196514	2010-07-24T13:44:38Z	53.3663709833	-2.2700764333	98503
196514	2010-07-24T13:44:26Z	53.3674087524	-2.2783813477	1043431
196514	2010-07-24T13:44:08Z	53.3675663377	-2.278631763	881734
196514	2010-07-24T13:43:18Z	53.3679640626	-2.2792943689	207763
196514	2010-07-24T13:41:10Z	53.364905	-2.270824	1042822

Check-in dataset sample

GeoFips	GeoName	LineCode	Description	2005:Q1	2005:Q2	2005:Q3	2005:Q4
00000	United States	1	Real GDP (millions of chained 2012 dollars)	14,767,846.0	14,839,707.0	14,956,291.0	15,041,232
00000	United States	2	Chain-type quantity indexes for real GDP	90.857	91.299	92.016	92.5
00000	United States	3	Current-dollar GDP (millions of current dollars)	12,767,286.0	12,922,656.0	13,142,642.0	13,324,204
01000	Alabama	1	Real GDP (millions of chained 2012 dollars)	182,600.4	184,625.7	184,134.9	186,116
01000	Alabama	2	Chain-type quantity indexes for real GDP	96.489	97.559	97.299	98.3
01000	Alabama	3	Current-dollar GDP (millions of current dollars)	155,702.7	158,097.7	159,237.5	162,349
02000	Alaska	1	Real GDP (millions of chained 2012 dollars)	45,176.1	45,776.9	45,501.7	46,173
02000	Alaska	2	Chain-type quantity indexes for real GDP	77.511	78.542	78.069	79.2
02000	Alaska	3	Current-dollar GDP (millions of current dollars)	37,792.5	39,198.7	40,905.2	43,529
04000	Arizona	1	Real GDP (millions of chained 2012 dollars)	256,521.4	260,884.6	265,619.7	266,278
04000	Arizona	2	Chain-type quantity indexes for real GDP	94.504	96.111	97.856	98.0
04000	Arizona	3	Current-dollar GDP (millions of current dollars)	220,647.6	225,552.6	231,487.1	233,976
05000	Arkansas	1	Real GDP (millions of chained 2012 dollars)	104,693.3	105,271.5	105,872.6	108,082
05000	Arkansas	2	Chain-type quantity indexes for real GDP	96.499	97.032	97.586	99.6
05000	Arkansas	3	Current-dollar GDP (millions of current dollars)	89,068.5	89,829.1	91,062.3	93,591

Financial dataset sample

Figure 2 Data samples

Together there are tens of millions of check-in data rows each with the longitude and latitude of the locations where the check-ins happened.

3.1.4 Financial dataset

The financial dataset is collected from the U.S. Bureau of Economic Analysis which contains the GDP and personal income of each state for each quarter over a long period of time that can cover all our datasets mentioned above.

3.2 Data Ingestion

In this section, we show you the data ingestion we did to get a clearer dataset for future analysis. There are some major challenges when doing the ingestion. First, the size of some dataset is too large, and there are not only posts in the US but all over the world. Second, some of the data, for example, check-in data contains only the latitude and longitude of the posts but not the actual state of that post, so there needs to be some transformation before we can do the analysis.

As for the social media dataset ingestion, we used Map-Reduce program to do the cleaning and extract the information we need. What we did is first extract the country information, if the country is missing or the country is not the United States, we excluded the data and went to the next one. We then extract the state information and time of the posts and combine them as the key to the Mapper output. The value of the Mapper output is set to 1 as normal counting program. In Reducer, we just added up the count for the same key to get the number of posts per month for each state. To notice that there are some datasets, like Gowalla dataset, without the attributes of state. For those datasets, we use the longitude and latitude data to get the state and country of the posts which we'll discuss later. The pseudo code of the Mapper program is shown in Figure 3.

As for the financial dataset, we use similar algorithms as described in Figure 3. The only difference is that the value now is the GDP. We only have GDP for each quarter, so we assumed the GDP for a month is the average GDP of that quarter.

Finally, we used Hive to combine the Posts dataset and GDP datasets. We used Join-operation to combine both datasets on key which is of the format

$$State, YYYYMM$$

so that we could do OLS regression to analyze the data and draw the figure that represents the relationship between economic status and the number of posts. We'll further discuss the evaluation method and process in Evaluation section.

3.2.1 Latitude and Longitude to State

To resolve the problems that some of the data contains only latitude and longitude data instead of the accurate position where the posts happened. We adopted the Google Reversed Geocoding (GRG) service to achieve our goals. The GRG provides an API that takes the latitude and longitude data and return a JSON object that contains the detailed

```
Mapper:
    posts = parse(value); // the data is parsed to json format
    country = posts['country];
    if country == 'US':
        if "state" in posts:
            state = posts['state'];
        else:
            state = get_state_from_latlong(latitude, longitude);
        time = posts['date'];
        time = reconstruct(time); // format the time to YYYYMM
        key = state + "," + time;
        value = 1;
        context.write(key, value);
```

Figure 3 pseudo code of the Mapper

```
X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,Texas,X,
X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,Texas,Texas,
X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,Texas,Texas,
X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,Texas,Texas,Te
X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,Texas,Texas,Texas,Texa
X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,X,Texas,Texas,Texas,Texas,
X,X,X,X,X,X,X,X,X,X,X,X,Arizona,Arizona,Arizona,Arizona,New Mexi
X,X,X,X,X,X,X,X,California,California,California,Arizona,Arizona
X,X,X,X,X,California,California,California,California,California
X,X,X,X,X,California,California,California,California,California
X,X,X,X,California,California,California,California,California,C
X,X,X,California,California,California,California,California,Nev
X,X,California,California,California,California,California,Nevad
X,X,California,California,California,Nevada,Nevada,Nevada,Nevada
X,California,California,California,California,California,Nevada,
X,California,California,California,California,California,Nevada,
```

Figure 4 Matrix that maps lat-long to state

location information from where we can extract the country and state information we need. However, there are couple of fatal problems with this approach. First, the GRG service is not cheap, especially when the requests are large. In our cases, there are tens of millions of requests need to be made which would costs thousands of dollars that is unaffordable and unnecessary. Another issue is that to use the API in HPC, we need to make HTTP requests to the Google API and wait for the response which would hugely slow the Map-Reduce process. As an example, using Map-Reduce program to process 10 million rows of data with state information would take around 1 minutes, however if we include the HTTP requests procedure, the time to complete the same task would be around 2 days which would also largely depend on the API's performance. These factors make the approach infeasible.

Therefore, we came up with another method to get the state and country of the posts. Since we only need the state and country information instead of the detailed street information, what we did is build a large matrix wrapping around the United States mainland. The abscissa of the matrix is longitude while the ordinate of the matrix is latitude. We used a granularity of 1 longitude and latitude. Then for each entry in the matrix, we used the GRG to get the state of that entry and put that into the entry. After we have done the process, we got a matrix that could map the latitude and longitude data to the state of the United States with only thousands of API requests to GRG. A part of the matrix is shown in Figure 4, 'X' means that entry isn't part of an US state. With the matrix, we can do the Map-Reduce program locally without issuing any HTTP requests which assures the process to be reliable and fast.

3.3 Analysis

We used OLS regression to analyze the relationship between number of posts and GDP. After that, we used some fitting methods to draw the relationship for each state. We also make analysis about whether we can predict a state's economic status using the number of social media posts and vice versa. We'll discuss the analysis method, results, and evaluation in detail in the next section.

4 Result

4.1 Overall Analysis

For each social media dataset, we will use OLS regression to detect the relationship between GDP and number of posts or logins in each month of each state. The results are examined with standard and t-test. In this analysis, we want to observe the relationship between social media activities and GDP across the states.

Also, we have separate regression on dataset before and after COVID-19.

4.1.1 Before COVID-19

For the datasets before COVID-19, the login data of the combination of Gowalla and Brightkite and post data of Instagram are separately regressed with GDP of each month and state. We get the result that $Login = 0.014 * GDP$ (Figure 5) for Gowalla and Brightkite, and $posts = 0.0171 * GDP$ (Figure 6) for Instagram. They both shows a positive linear relationship, which means a higher GDP will result in a higher frequency of activities on social medias.

Both of the coefficients are examined with t-test and get the result that $t > 30$ and $P(|T| \geq |t|) < 0.001$, which shows a very high statistical confidence on the positive linear relationship.

Also, both of the coefficients have a standard error less than 0.0001, which shows that the real distribution is very close to our regression line.

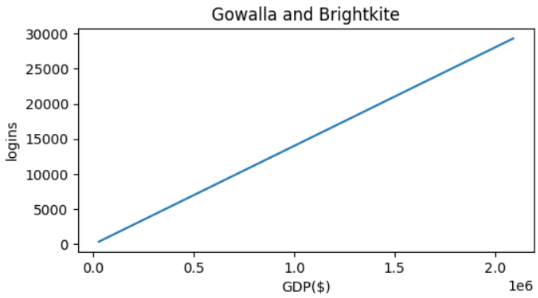

Figure 5: The regression line between GDP and number of logins of Gowalla and Brightkite.

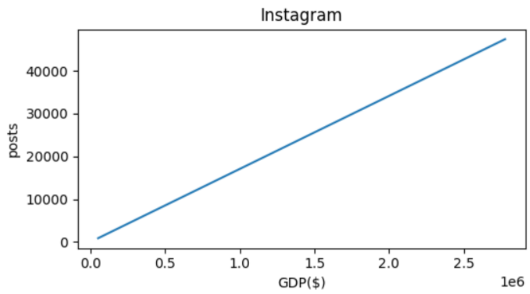

Figure 6: The regression line between GDP and number of posts of Instagram.

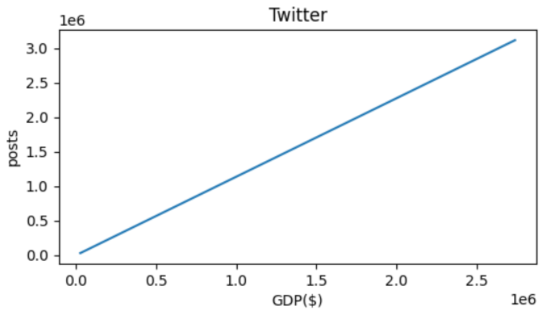

Figure 7: The regression line between GDP and number of posts of Twitter.

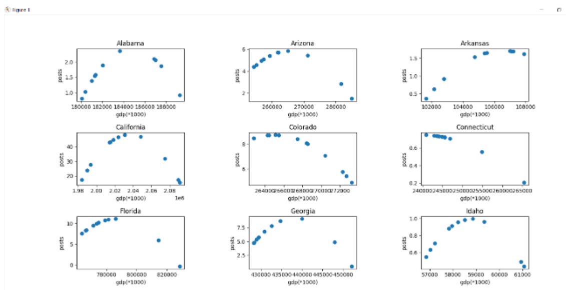

Figure 8: The regression line of 9 unbiasedly selected sample states

4.1.2 During COVID-19

The posts data of Twitter for 3 months and 50 states during COVID-19 is regressed with GDP. We get the result that $No.Posts = 1.1367 * GDP$ (Figure 7), which shows a positive linear relationship between GDP and number of twitter posts.

The coefficient is examined with t-test and get the result of $t = 17.556$ and $P(|T| \geq |t|) < 0.001$, which shows a very high statistical confidence on the positive linear relationship. However, the t-value is significantly lower than the standard error of regressions before COVID-19. This means that the COVID-19 will decrease our confidence of the positive linear relationship between GDP and social media activities. But even with the effect of COVID-19, we still have enough confidence to our regression line.

Also, the coefficient has a standard error of 0.065 which is very low relative to the coefficient. This shows that the real distribution is very close to our regression line.

4.2 Monthly Analysis

For each state, we also separately regress the social media activity data with the GDP of that state in each month. In this analysis, we want to observe the relationship between social media activities and economic situation.

For most state, we got a convex curve like Figure 8. Although a few states have a low t-value which means the result is insignificant, we still have enough confidence for most of states that the social media activities are less frequent when the economic situation is at a relatively low level or high level. This is consistent even during the COVID-19.

This analysis shows a different curve with overall analysis. We think there are 2 reasons for this. 1) The high or low frequency of social media activities in this analysis is relative to each state, so it has less effect on overall activities. 2) The up or down of economic situation is not in same frequency. We could observe from Figure 8 that the low or medium economic situation is more frequent than high economic situation. In this way, the part of the curve which is shifting up could be explained by coefficient of the overall regression line, while the part of shifting down curve could be explained by the variance of the regression line.

5 Limitation and Extension

5.1 Limitations

Even though we have come up with logical results, there are still some limitations may influence the accuracy of our project.

Firstly, we did not find a perfect way to deal with the different granularity of GDP dataset and Post datasets. In our GDP dataset, the unit is GDP per quarter. However, in our Post datasets, the unit is number of posts per month. We tried to convert the unit from per month to per quarter in our Post datasets for accuracy concerns, but it is impossible since in some of the post data, month are not consecutive so that it cannot be grouped into a quarter. Based on this situation, we split the GDP per quarter into three equal-value GDP per month. All our following analysis will be based on data of each month. In our final dataset, when there is fluctuation in the number of posts per month, the GDP of three consecutive month will remain the same. This prevents us to get an accurate relationship between the trends of number of posts and the trends of GDP. This will not influence the final result too much, but we will lose some details as a consequence.

Secondly, there will be many other factors influence the GDP. For example, the number of industrials and companies will have large influence on the local GDP. However, the number of posts have little relationship with those companies because we find that most of the posts are published by individual users. To make our project more accurate and useful, we need to take other factors like number of companies into consideration.

5.2 Extensions

Our project figures out the number of posts and the GDP of each state. This can be used to predict the GDP of current month. In the first several days of each month, we can calculate how many posts are there is several social medias. By scaling the data to one month, we can estimate what is the total number of posts of that month. So that we can estimate what is the expected GDP of that month. This prediction will be meaningful to government to predict the financial situation in different state. Investors can also use this prediction to see what the trends of the market is and help them make decisions.

To achieve this, we will need another model, which will calculate the distribution of posts each day of a month. With this model, we can predict the total number of posts with only data of several days in a month.

6 Conclusion

In our project, we find that there is relationship between economic situation and number of posts in each state in the United States. The relationship follows a convex curve: When GDP is too low or too high, the number of posts will decrease rapidly. When it is in the middle, the number of posts will reach its peak.

We analyze the reason for this as: when the economic situation is not good, people do not have so much entertainment consumption to post on social networks. Instead, they will look for some way to survive under this terrible situation. After that, when the economic situation improves, people have more activities and post it on social media to show off. In this time, people do not have worries about their financial status, neither will day have enthusiasm to explore new ways to make money since there is no significant chances to achieve it. Finally, when the GDP is very high, people may disdain to continue to find a sense of presence on social networks, it’s a waste of time. they may have higher spiritual pursuits and reduce posting on social media. Or they are finding methods to earn money in such an economic prosperity.

References

[1] U.S. Bureau of Economic Analysis, Gross Domestic Product [GDP], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GDP, May 8, 2022.

[2] Chen Ming, Zhu Xueshuai and Jiang Qian. 2019. *Challenges to Traditional Statistics in the Age of Big Data.*

[3] Agustín Indaco. 2020. From twitter to GDP: *Estimating economic activity from social media,volume 85.* Regional Science and Urban Economics

[4] U.S. Census Bureau, Real Median Household Income in the United States [MEHOINUSA672N], retrieved from FRED, Federal Reserve Bank of St. Louis.https://fred.stlouisfed.org/series/MEHOINUSA672N, May 8, 2022.

E. Cho, S. A. Myers, J. Leskovec. 2011. *Friendship and Mobility: Friendship and Mobility: User Movement in Location-Based Social Networks ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Umair Qazi, Muhammad Imran, & Ferda Ofli. (2020). *GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information [Data set]*. In ACM SIGSPATIAL Special (V.1.0.0, Vol. 12, Number 1, pp. 6–15). Zenodo. *https://doi.org/10.5281/zenodo.3878599*

X-Byte. 2022. *Instagram post details scraper - extract instagram posts, public profiles, reels, hashtags or locations [Data set].*

Bureau of Economic Analysis U.S. Department of Commerce. *GDP and Personal Income [Data set]*