Appendix A – Master Project Cover Page

PROJECT TITLE

By

Kunal Jangam

A Research Project Submitted in
Partial Fulfillment of the
Requirements for the
Degree of Master of
Science in Statistics
Specialization in Financial Statistics and Risk Management (or Data Science for MSDS)
At
Rutgers, the State University of New Jersey

December 2022

PROJECT ABSTRACT:

In this project, I use the rtweet package to perform data analysis on tweets, split into two datasets, one containing tweets with the keyword “pandemic” and the other containing tweets with the keyword “covid”.

Through this project I want to understand the public perception of covid-19 and the pandemic and see if people still take preventative measures against covid-19 or if they have started to live as if covid-19 is no longer a threat. I also want to explore twitter data and see how it compares to the data I am used to using.

ACKNOWLEDGEMENT

I would like to thank Jinwoo Lee for working with me to complete this project. I would also like to thank Professor Tirthankar Dasgupta for training me in data wrangling, and for supervising this project.

TABLE OF CONTENTS

TABLES

Pandemic tweets popular words

word <chr>	n <int>
pandemic	9602
https	5662
covid	1583
people	1075
amp	945
home	891
health	658
williamchurch	648
time	620
started	535

Pandemic tweets popular bigrams

bigram <chr>	n <int>
pandemic https	453
bitcoin trading	324
home isn	324
isn easy	324
met williamchurch	324
started bitcoin	324
post pandemic	240
covid pandemic	234
global pandemic	190
public health	173

Pandemic tweets popular words revisited

All Data

word <chr>	n <int>
pandemic	9602
https	5662
covid	1583
people	1075
amp	945
home	891
health	658
williamchurch	648
time	620
started	535

>10 Likes

word <chr>	n <int>
pandemic	407
https	293
covid	67
amp	56
people	42
health	35
time	30
don	24
government	22
world	21

Pandemic tweets popular bigrams revisited

All Data

bigram <chr>	n <int>
pandemic https	453
bitcoin trading	324
home isn	324
isn easy	324
met williamchurch	324
started bitcoin	324
post pandemic	240
covid pandemic	234
global pandemic	190
public health	173

>10 Likes

bigram <chr>	n <int>
pandemic https	26
post pandemic	13
global pandemic	10
covid pandemic	9
pre pandemic	7
public health	7
health care	6
matt hancock	6
pandemic amp	6
climate change	5

Covid tweets popular words

word <chr>	n <int>
covid	10202
https	5073
de	3003
la	1611
en	1084
el	950
le	684
people	677
amp	548
por	525

Covid tweets popular words

All Data

word <chr>	n <int>
covid	10202
https	5073
people	677
amp	548
don	395
deaths	380
time	321
mask	315
vaccine	302
health	290

>10 Likes

word <chr>	n <int>
covid	233
https	138
people	19
mask	14
amp	13
death	13
days	11
die	11
health	11
deaths	10

Covid tweets popular bigrams

All Data

bigram <chr>	n <int>
covid https	323
covid deaths	104
da covid	84
post covid	84
vargas llosa	84
covid test	81
covid vaccine	67
emerg sanit	60
mario vargas	56
oficializa fim	47

>10 Likes

bigram <chr>	n <int>
mask mandate	5
covid https	4
post covid	4
wear masks	4
covid deaths	3
evoluciona favorablemente	3
tested positive	3
vargas llosa	3
assina portaria	2
bad person	2

NRC sentiment Analysis tables

Joy

Pandemic

word <chr>	n <int>
money	473
love	131
food	115
hope	113
pay	109
safe	89
happy	88
found	82
share	79
deal	70

Covid

word <chr>	n <int>
hope	96
money	68
safe	65
love	61
pay	61
feeling	59
food	55
god	54
found	47
true	46

Sadness

Pandemic

word <chr>	n <int>
pandemic	9602
lost	472
death	110
bad	107
worse	77
sick	71
hospital	67
die	58
deadly	56
disease	53

Covid

word <chr>	n <int>
pandemic	258
death	230
die	206
hospital	171
sick	113
disease	102
bad	96
negative	88
lost	78
cancer	67

Anger

Pandemic

word <chr>	n <int>
money	473
hit	145
death	110
bad	107
demand	89
shit	67
fight	61
deadly	56
disease	53
dying	53

Covid

word <chr>	n <int>
death	230
disease	102
bad	96
money	68
cancer	67
feeling	59
dying	57
hit	55
sin	54
shit	52

Fear

Pandemic

word <chr>	n <int>
pandemic	9602
government	269
inflation	187
war	177
change	134
death	110
bad	107
flu	90
risk	84
watch	78

Covid

word <chr>	n <int>
pandemic	258
death	230
die	206
hospital	171
risk	170
flu	135
government	132
disease	102
infection	97
bad	96

Trust

Pandemic

word <chr>	n <int>
money	473
don	430
provide	360
real	162
economy	159
school	149
policy	121
food	115
hope	113
pay	109

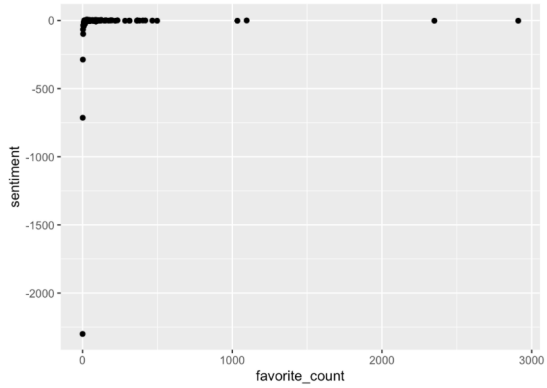
Covid

word <chr>	n <int>
don	395
hospital	171
wear	126
county	101
school	97
hope	96
leading	91
team	78
policy	71
money	68

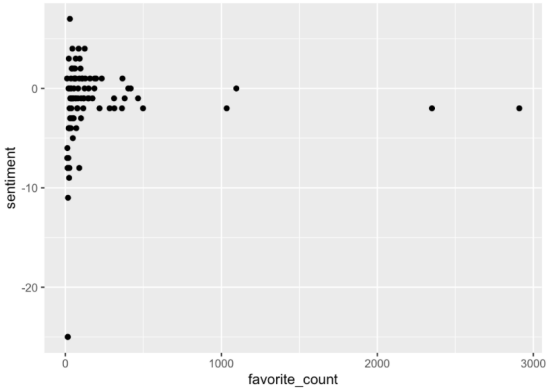
GRAPHS: plotting sentiment against favorite account

Pandemic sentiment analysis

All Data

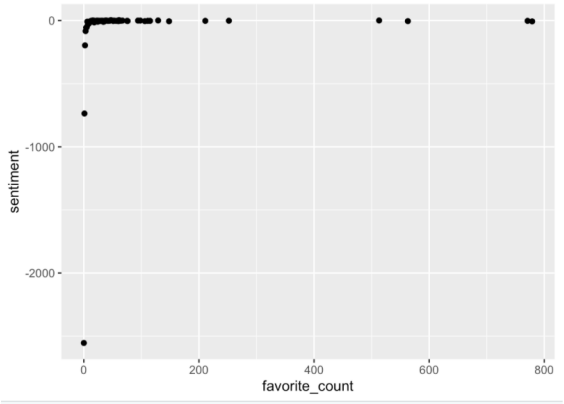


>10 Likes

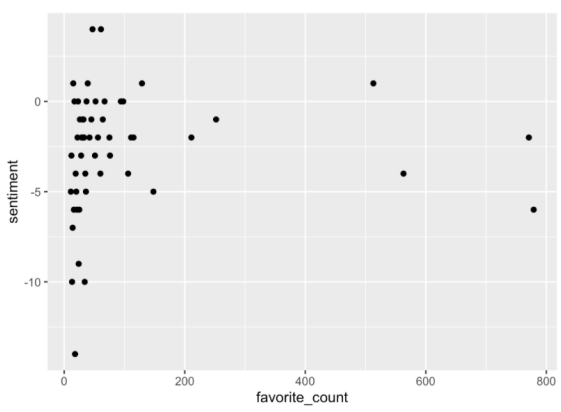


Covid sentiment analysis

All Data



>10 Likes



Tables

INTRODUCTION

In the year 2020, when the pandemic first began, many people proceeded with caution, taking measures such as wearing masks and quarantining themselves. Two years later, vaccines have been created to combat covid-19. However, covid-19 has not disappeared and is still infecting many people every day. One goal of this project is to determine if the public is still cautious of covid-19 and is still taking precautionary measures to stay safe, or if the public no longer sees the disease as a threat.

I will collect data from recent twitter posts using the rtweet package. Since I am using twitter data, the second goal of this project is to understand the differences twitter data has compared to data we are used to using.

In order to accomplish these goals, I will use R to perform data wrangling to generate meaningful tables and graphs. I will find the most popular words and bigrams for tweets talking about covid or the pandemic. I will perform sentiment analysis to compare and contrast tweets of joy, of sadness, of anger, of fear, and of trust, and to create plots to see which of these sentiments is common and popular on twitter. I will analyze this information to compare and contrast tweets that reference covid and tweets that reference the pandemic.

CHAPTER 1: Pandemic Tweets Popular Words and Bigrams

We start by extracting the top ten most popular words included in tweets that contain the word "pandemic". The result is displayed in the Pandemic tweets popular words table of page four of this document. Using this table, we can see that over half these tweets contained the string "https", meaning the tweet either contained a link to a website or article, presumably to provide evidence for their argument or to start a discussion, or the tweet referenced another tweet, meaning some conversation took place between users about the pandemic.

A peculiar string that appears in the table is williamchurch. It is unexpected for some unknown person's name to be so common among tweets referencing the pandemic. To understand the reason for this phenomenon, we look at the results of the top ten popular bigrams included in tweets that contain the word "pandemic", shown in the Pandemic tweets popular bigrams table of page four of this document. From this table we see that several bigrams, including "met williamchurch", "bitcoin trading", and "started bitcoin", all had the exact same number of occurrences. This implies that the exact same message about William Church and bitcoin has been posted hundreds of times. From this information we can conclude that William Church is a crypto investor who uses bots to spread his name around twitter. A quick search on twitter for the bio of William Church confirms this fact. Through this process, we were able to successfully analyze the data to understand the goals and methods of a twitter user.

To filter out these bot tweets from the data, we create tables only including results from tweets that have more than ten likes. The reason for this is that in general bot messages do not have many likes. As we can see in the revisited tables on pages four and five of this document, this filter was successful in removing the impact the bot messages had on the table. This process has shown that twitter has a problem with bots, and this problem has to be considered when analyzing recent posts.

CHAPTER 2: Covid Tweets Popular Words and Bigrams

Next, we analyze the popular words and bigrams from tweets that included the word “covid”. From the first Covid tweets popular words table on page six of this document, we observe that many of the top ten words are spanish stop words, such as el or la. This means that a substantial amount of the tweets collected were in spanish.

The pair Covid tweets popular words tables on page six of this document shows the results of removing the spanish stop words from the previous table. From these tables we can see that most of the tweets mentioning covid also mention masks, vaccines, death, and health. Similar to the pandemic tweets tables, links are in over half the tweets. On the other hand, most tweets that mentioned covid did not mention the word “pandemic”. We can conclude that, from the high frequency of words such as masks, vaccines, death, and health, that the people on twitter who are talking about covid take it seriously.

While the table for popular words does not explain the presence of spanish stop words in the dataset, the table for Covid tweets popular bigrams on page seven of this document does. The names Mario Vargas and Vargas Llosa are common among tweets talking about covid. As such, it is probable that Mario Vargas Llosa is a famous person who either recently contracted covid or passed away from covid. Fortunately, after searching through the news, we learn that Mario Vargas Llosa is a Peruvian born Spanish nobel prize winning novelist who contracted covid but is diagnosed to make a recovery. Through this process, we have successfully analyzed our tables to learn about recent occurrences.

From these results we can see that there are still many who consider covid a threat, and that tweets about covid can revolve around news of a famous person falling sick.

CHAPTER 3: Sentiment Analysis

Our last form of analysis was sentiment analysis. We created tables that contained popular words categorized by sentiment. These tables are on pages 8-10 of this document. These tables gave more insight on how the public views the pandemic and covid. The Fear table on page nine and the Trust table on page ten of this document were particularly interesting. From these tables, we see that for tweets talking about the pandemic, the words government, inflation, demand, change, money, and economy were popular, while for tweets talking about covid, the words disease, hospital, and deaths were popular. This most likely means that many people talking about the pandemic are complaining about how it is being handled by the government and how the economy is suffering because of inflation. They demand change from the government. On the other hand, tweets talking about covid may be arguing the opposite side, using the number of deaths and hospitalizations as evidence against removing covid restrictions.

We also compared sentiment to the popularity of a post, measured by the number of favorites a post has. Graphing the net positivity of a tweet against the favorite count of the tweet resulted in the graphs on page 11 of this document. From these graphs we can see that the most popular tweets had almost neutral positivity/negativity, with a slight trend towards negative posts being more popular than positive ones. In addition, some posts were extremely negative, but received very few likes and were removed from the graph by filtering out posts with less than ten likes.

CHAPTER 4: Conclusions

Collecting popular words and bigrams of tweets along with performing sentiment analysis provided interesting insight on how the public treats the covid pandemic two years after it began and how twitter posts are in general. We have learned that many people still consider covid to be a threat, and that there are many others who want to risk a return to normalcy for the sake of the economy. We have learned that twitter has a problem with bot accounts, and that twitter posts often follow recent news. We conclude that people talk about the pandemic in a different way from how they talk about covid.

CITATIONS:

Information about Mario Vargas Llosa is from

https://www.ndtv.com/world-news/nobel-laureate-mario-vargas-llosa-hospitalised-with-covid-2912436