CSC110 Project Report
DETERMINING THE CORRELATION BETWEEN THE IMPACT OF UNIVERSITIES SOCIAL MEDIA ANNOUNCEMENTS AND COVID CASES USING STATISTICAL ANALYSIS

Kurtis Law, Minh Le, Shaan Purewal, Hyun Bin Kim

December 14, 2021

Problem Description and Research Question

Research question: How do Ontario universities' COVID-19 announcements correlate to local cases?

As of October 29th 2021, nearly 40% of all COVID-19 cases in the country are linked to university-aged people (19-29 years old) (Public Health Agency of Canada). Throughout March and April 2021, outbreaks were declared individually at the University of Waterloo, Queen's University, Brock University and Western University (LeBel). The public health unit at Kingston Ontario said over 70% of the area's COVID-19 cases were linked to neighbourhoods around Queen's University (LeBel). Without a doubt, universities are a non-negligible factor in the country's number of COVID-19 cases.

Thankfully, universities each took to their own way of responding to this pandemic. On social media, universities each announce their way of fighting the pandemic: policy changes, resource distribution, switch to online lessons, etc. Since different universities had different plans in place at different times, we thought they would serve as interesting semi-isolated data points to cross compare COVID-related announcements to different changes in COVID-19 cases. Hence, using data, we plan to explore **how universities' COVID-19 announcements correlate to local cases.**

Note that the model will aggregate all COVID-19 related announcements and include them into our model, no matter their suggested importance. This is to simplify an otherwise extremely arduous process of sorting through thousands of social media posts each university has posted throughout the course of the pandemic. To adjust to this, we will create our own keyword-to-value dictionary that assigns each announcement a nominal value - we call impact score, as an impartial way to judge the announcement's impact. For example, an announcement mentioning 'online school' will be judged as being more impactful than an announcement that mentions 'self check-ins'. A complete dictionary mapping keywords to their impact score can be found in our Python programs. This list is created based on the major policies, mandates, and other rules and advice that universities have implemented so far in the pandemic.

At the end of our project, we should arrive at the correlation between the impact score and the change in local COVID-19 cases, hence answering the research question.

Datasets Description

Announcements data is collected from the Twitter of five Ontario universities. Local COVID-19 data is collected from a variety of city/regional websites.

The universities:

1. Brock University
2. Queen's University
3. University of Toronto (St. George Campus)
4. University of Waterloo
5. Western University

COVID-19 public announcements (social media):

Twitter data collected using the library Twint, then stored as csv files consisting of a variety of tweets' information including date, universities' names, and the tweets' content. The files downloaded will be stored in the folder 'Datasets' and named as following (sorted alphabetically):

1. Tweets_Dataset_BrockUniversity.csv
2. Tweets_Dataset_queensu.csv
3. Tweets_Dataset_UofT.csv
4. Tweets_Dataset_UWaterloo.csv
5. Tweets_Dataset_WesternU.csv

After the files are stored, only the 'date' and the 'tweets' columns will be extracted to use in further analysis.

COVID-19 cases (location based):

1. Collected OntarioCovid_Dataset.csv from the data.ontario.ca.
2. Use built-in library `requests.py` to download 'Confirmed positive cases in COVID-19 in Ontario'.
3. Stored as csv file (mapping date, location in latitude longitude coordinate form)

Computational Overview

Our project composes of four parts:

1. Data Collection

   All datasets involving recorded tweets were sourced using the `twint` library. For specific use, files setup.py and Twint_Scrape.py include all code/documentation required to collect tweets from Twitter. In summary, `configure_url()` and `configure_token()` functions alter configuration files within the `twint` package to minimize errors and limitations (further discussed below), `collect_tweets()` is then called given a Twitter username. Utilizing `twint`, we pass through some configuration such as storing method and weeks since tweets, we are able to collect all tweets within a specified time period (since Mar 20, 2020 - present).

   Our COVID-19 data had been sourced from the Ontario Data Catalogue, a public and free government data site. This dataset is automatically downloaded in our program. To access the data, you may directly download or visit the resource page for more detailed instructions. If the user has failed to download the COVID-19 dataset, `download_covid_data()` (within `setup.py`) will download the dataset from the Ontario Data Catalogue with the `requests` (built-in) library. If the dataset isn't detected, a GET request of the direct download link will download the data file to the Datasets directory. To conclude, all data is stored in respective csv files within the 'Datasets' directory after initial setup.

2. Data Processing

   The keyword-impact score dictionary used throughout the project is created based on available reports on the most effective methods and policies to prevent the spread of COVID-19 (Wibbens, Phebo D., et al). For example, 'mask' is assigned an impact score of 4/10, while 'vaccinations' is assigned a score of 8/10, based on their effectiveness against the virus.

   (a) Filter Twitter files

   The data collected from the five universities' Twitter profiles will be passed as csv files into the function `filter_uni_tweets()`. Unwanted attributes of the tweet such as hashtags, place, and id are removed. Tweets unrelated to COVID are also removed. Afterwards, the impact score of each tweet is calculated and tweets are grouped into weeks using modular arithmetic - adding up to the cumulative impact score of the week (if there are multiple COVID-related tweets per week). Grouping tweets into weeks instead of days reduce anomalies in the data and would produce a better regression estimate in the later stage of the project. Finally, a csv file is created that maps the week to its corresponding impact score, where the week is the number of weeks since March 20th, 2020 - the first day in the range of the data collected.

   (b) Filter Ontario COVID Dataset

   We want to determine the number COVID cases per week that are related to the five universities. The radius of the affected area is set to be 5km, with the center of the circle being the coordinate of the campus. Since the Ontario COVID dataset contains specific location coordinates of the infected cases, we can then look for cases detected within the calculated area. The function `filter_ontario_covid_data()` removes unwanted columns in the Ontario COVID dataset and maps universities to their related COVID cases. Within this function, helper functions determine whether COVID cases are within the radius specified. For example, the function `within_range()` returns whether two coordinates are within a specified distance, while the function `haversine()` determines the distance between two points on a sphere, thereby finding the distance between two coordinates on Earth. In the end, a csv file is created for each university, in which the date is mapped to the COVID cases per week related to that university.

3. Data Visualization

   At this stage, for each of the five universities, we have two final csv files: one mapping weeks to the cumulative impact score per week of that universities' tweets, and one mapping weeks to COVID cases related to that university. These csv files are ready to be plotted.

Using `matplotlib`, as an example, we plot the data for the University of Toronto as follows:

(a) Cumulative impact score vs. Weeks since 3/20/2020

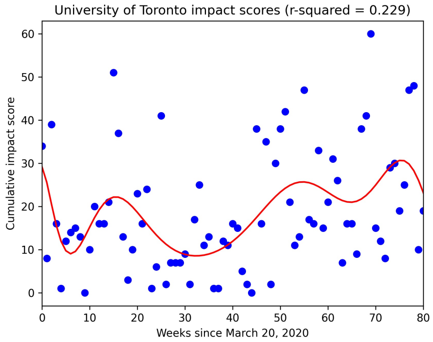


Figure 1: Impact score vs. Weeks

(b) Local COVID cases vs. Weeks since 3/20/2020

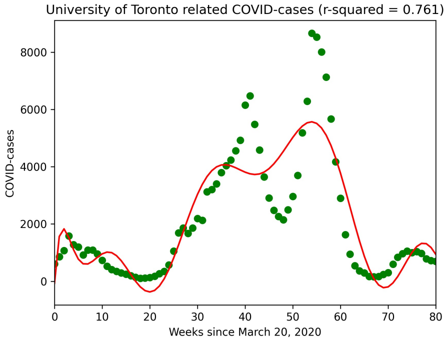


Figure 2: COVID cases vs. Weeks

4. Data Analysis

First, we determine how accurately the trendlines calculated by numpy represents the data in the scatterplots by calculating the coefficient of determination r-squared (r^2). In our case, for the plot between weeks and COVID cases, r-squared represents the percentage of variance in COVID cases (the dependent variable) that can be explained by week (the independent variable) through the calculated regression model (the trendline). The closer this value is to 1, the more accurate the regression model is, though a reliable value would be greater than 0.4. We also analyze the residuals of the data points about the regression line - this is the difference between the predicted data and the actual data. In an accurate regression model, we expect to see an "evenly distributed" pattern of residuals, meaning that data points are randomly positioned above and below the trendline.

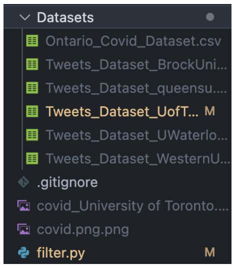
After analyzing the accuracy of the model, we proceed to interpret these trendlines and compare the two - the trendline for impact score vs. week and the trendline for COVID case vs. week.

Our hypothesis: **impact score and COVID cases has an inversely proportional relationship**.

As universities implement stricter measures and policies to protect themselves from COVID-19, as reflected by the increase in impact score from their social media announcements, we expect to see a decrease in local COVID cases. It is important to note that correlation $\neq$ causation. This project only aims to determine the correlation between the two variables and assumes no conclusions on whether the change in one variable affects the other. However, further studies involving isolating parameters can help us arrive closer at determining if there exists a causation relationship, which will be suggested in the Discussion section below.

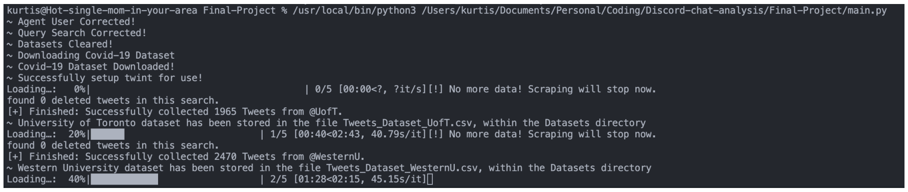
Instructions to Run the Program

1. Download all the files in MarkUs. Place them into one file. Make sure they are all under one directory.
2. Check the `requirements.txt` file, install all the necessary libraries.
3. Run `main.py`
    - At first, the `setup.py` will run. The purpose of this file is to correct the twint library so that the tester will not face any possible errors while scraping data from twitter.
    - Secondly, the program will proceed to scrape all tweets from the universities and store the data in each as CSV files under the directory named 'Datasets'. These files will include all tweets from each universities' Twitter accounts since March 20th, 2020.
        - Though this functions makes it so that the TA does not require to manually download any of the raw files, in case any of the files fails to download, we will link 'Datasets' as per instrcuted by the handout. Claim ID: 5qkRTivVXv2F4YMW, Claim Passcode: UWoXtSxa4ChxUm34



```
Datasets
    Ontario_Covid_Dataset.csv
    Tweets_Dataset_BrockUni...
    Tweets_Dataset_queensu....
    Tweets_Dataset_UofT...  M
    Tweets_Dataset_UWaterlo...
    Tweets_Dataset_WesternU...
.gitignore
covid_University of Toronto....
covid.png.png
filter.py                   M
```

- Thirdly, the file will call `compile_universities`, which is a function from `filter.py` This python file achieves two important filtering of the data for clean visualization: filter each universities tweets and assign with impact scores, and filter the Ontario COVID dataset. For each task completed, the terminal will update the status of completion.



```
kurtis@Hot-single-mom-in-your-area Final-Project % /usr/local/bin/python3 /Users/kurtis/Documents/Personal/Coding/Discord-chat-analysis/Final-Project/main.py
~ Agent User Corrected!
~ Query Search Corrected!
~ Datasets Cleared!
~ Downloading Covid-19 Dataset
~ Covid-19 Dataset Downloaded!
~ Successfully setup twint for use!
Loading…:   0%|                                      | 0/5 [00:00<?, ?it/s][!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
[+] Finished: Successfully collected 1965 Tweets from @UofT.
~ University of Toronto dataset has been stored in the file Tweets_Dataset_UofT.csv, within the Datasets directory
Loading…:  20%|██████                        | 1/5 [00:40<02:43, 40.79s/it][!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
[+] Finished: Successfully collected 2470 Tweets from @WesternU.
~ Western University dataset has been stored in the file Tweets_Dataset_WesternU.csv, within the Datasets directory
Loading…:  40%|████████████                  | 2/5 [01:28<02:15, 45.15s/it]▯
```

- Finally, with the filtered data, the visualization python file is ran, running both the functions `plot_impact` and `plot_covid`. These functions will show all the COVID-19 cases that were near each universities, counting by each week starting from March 20th.

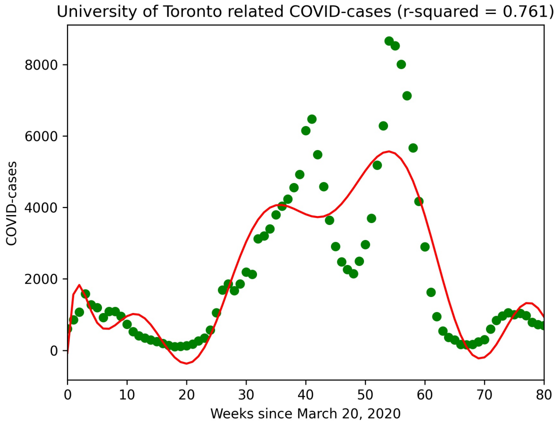


Figure 3: Example output showing COVID cases data for the University of Toronto

Changes Between Project Proposal and Final Submission

- In our project proposal, our TA has commented that we should specify what type of analysis we are going to use for our project. We have concluded to use statistical analysis to analyse the impact of tweets to the the number of COVID-19 cases near the university.

- As our TA pointed out, we modified to limit the collection of data to only Twitter instead of multiple other platforms. Collecting data across multiple platforms would not make much difference in our data, since each university uploads the similar posts for each platform, making it tedious to remove any duplicates.

- Instead of focusing on data collection, we have changed our focus on the computation and visualization. We have previously stated in our project proposal that the data will be reassigned with a lower-case version of themselves. However, converting the tweets to lowercase caused a lot of problems, as the emojis or special texts used in most tweets were in convertible. Instead, we have decided to delete such a process, and add more keywords in order to encompass possible variants of the word.

Discussion

1. Analyzing Accuracy of Regression Models

   As explained in our Computational Overview - Data Analysis, a statistic to analyze the accuracy of the regression model is r-squared - the percentage of observations that can be explained by the model. The formula for calculating r-squared is included in `visualisation.py`, and the values are calculated as follow:

   COVID CASES VS. WEEKS

   Brock University: 0.539
   Queen's University: 0.942
   University of Toronto (St. George Campus): 0.761
   University of Waterloo: 0.774
   Western University: 0.593

   IMPACT SCORES VS. WEEKS

   Brock University: 0.472
   Queen's University: 0.515
   University of Toronto (St. George Campus): 0.229
   University of Waterloo: 0.320
   Western University: 0.473

   Almost all of these values are above 0.4, indicating a moderately strong correlation. For the lower values (such as UofT and UWaterloo Impact Scores), assessing the regression models shows that the observations are distributed evenly above and below the trendline, which means that the difference between the regression line and the actual data are random. Thus, the regressions are fairly accurate and can be used for further analysis.

2. Interpreting Trendlines and Answering the Research Question

   We now have two relationships: one between impact score and days, and one between local COVID cases and days. Because in both of the relationships, the independent variable (days since 3/20/2020) has the same range, we can superimpose the trendlines on top of each other and analyze the relationship between impact score and local COVID cases.

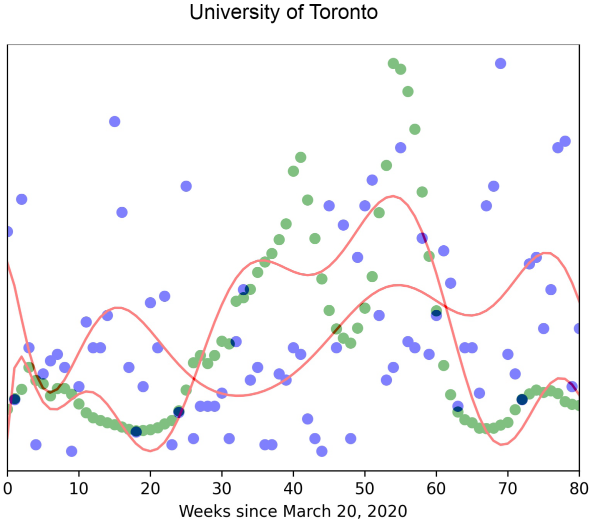


Figure 4: Superimposed regression models for the University of Toronto

   The figure above clearly shows that there exists an inverse proportional relationship between impact score and local COVID cases. For example, between week 10 and 20, COVID cases are at a trough while the impact score trend is at a crest. This time period corresponds to mid-July to mid-August, the period after the first wave of cases in Ontario.

This inverse relationship is also shown in all other universities, as shown in the superimposed graphs below:

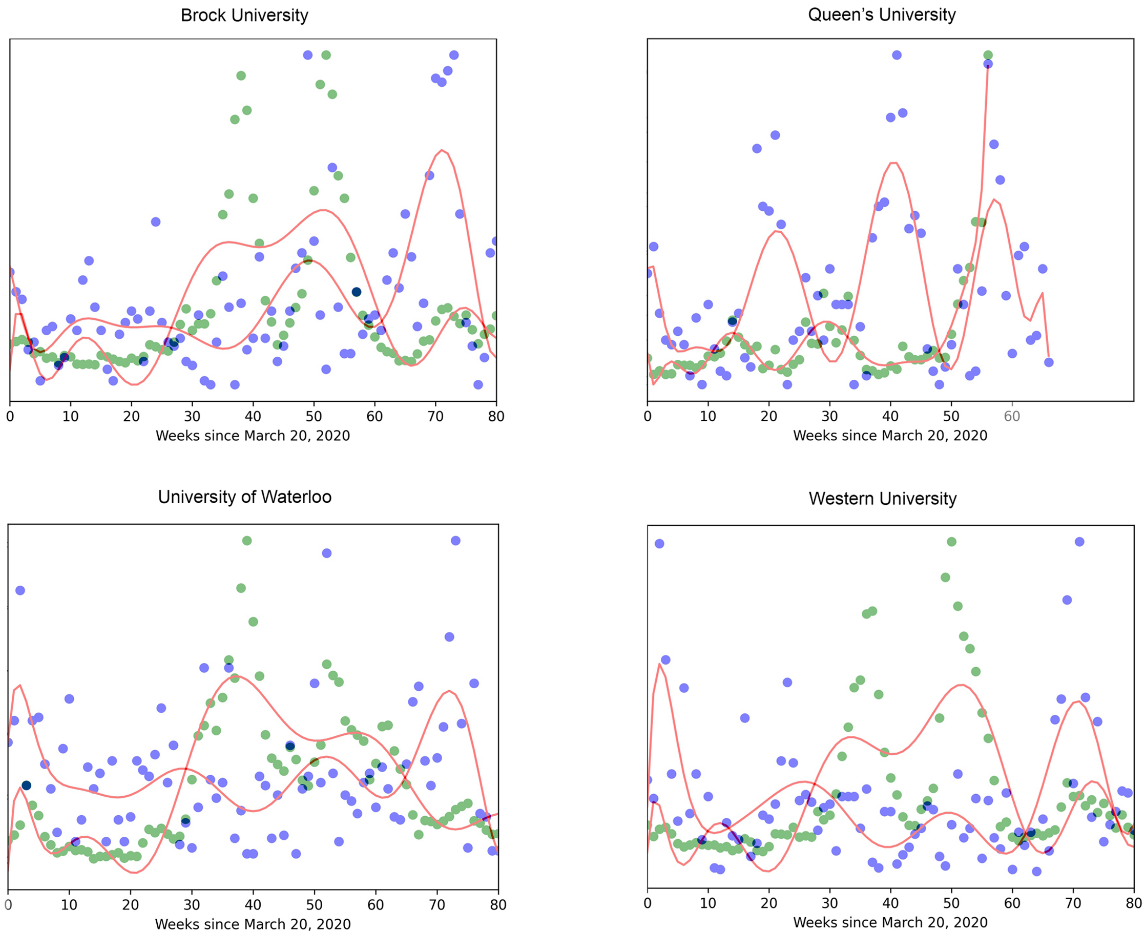


Figure 5: Superimposed regression models for Brock University, Queen's University, University of Waterloo, and Western University

Hence, we arrived at the conclusion that **there exists an inversely proportional relationship between universities' impacts on social media and the number of local COVID cases**. In other words, as universities announce stricter COVID-19 measures, which is reflected by the increased impact scores, the infection rates in the neighborhood around that university decreases.

3. Assumptions in the Analysis

This statistical analysis based itself on the assumptions that:

(a) Members of universities maintain complete compliance with the policies

(b) Announcements not made on social media will not be taken into consideration (as it is unavailable for scraping)

(c) Universities are a significant influence on the spread of COVID-19 cases in the surrounding area

(d) Symptoms take 14 days to appear after exposure to COVID-19

(e) Announcements take exactly 14 days to make an impact on the region's cases

4. Limitations and Possible Errors

A limitation is that we have limited the keywords list to a certain amount, hence not accounting for every keyword that could possibly impact people's awareness. We have also limited the five universities to be within the state of Ontario, and the lack of data could possibly skew the result. Furthermore, the value of impact scare we have assigned for each keyword could possibly not reflect our assumption, as there was no definitive metric for assigning these impact scores. As of result, the graph could contain anomalies, creating noise in the graph, lowering the accuracy. The effects of these possible errors have been minimized through analyzing the accuracy of the regression models.

5. Moving Forward: What's Left to Answer and Further Studies

   As previously stated, our project does not have the sufficient tools and models to conclude on whether the impacts of universities' tweets cause the change in COVID cases or vice versa. Perhaps the spikes in cases urges universities to announce stricter measures, or perhaps measures actually lowered the cases. We can only conclude that there is an inversely proportional relationship. We can proceed to isolate parameters affecting the two variables such as age group, commuters vs. in-res students, local outbreaks outside the university, population density, and others.

References

City of Toronto. “Covid-19: Case Counts.” City of Toronto, 23 Sept. 2021, https://www.toronto.ca/home/covid-19/covid-19-pandemic-data/covid-19-weekday-status-of-cases-data/.

LeBel, Jacquelyn. “Why Are Covid-19 Cases Quickly Rising among Canadian University Students?” Global News, Global News, 15 Apr. 2021, https://globalnews.ca/news/7757102/covid-19-canadian-university-students/.

“Matplotlib.” License - Matplotlib 3.4.3 Documentation, https://matplotlib.org/stable/users/license.html.

Public Health Agency of Canada. “Covid-19 Daily Epidemiology Update.” Canada.ca, 28 May 2021, https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html.

Twintproject. “Home · Twintproject/Twint Wiki.” GitHub, https://github.com/twintproject/twint/wiki.

“What Is Numpy?” What Is NumPy? - NumPy v1.21 Manual, https://numpy.org/doc/stable/user/whatisnumpy.html.

Wibbens, Phebo D., et al. “Which Covid Policies Are Most Effective? A Bayesian Analysis of Covid-19 by Jurisdiction.” PLOS ONE, vol. 15, no. 12, 2020, https://doi.org/10.1371/journal.pone.0244177.