# Project Report

# IDS 515: Information Systems Strategy and Policy

Twitter Data Analysis for Illinois Hospitals

Professor Ranganathan Chandrasekaran

Date: December 11, 2020

## ABSTRACT

The American Hospital Association is conducting a country wide analysis to determine the COVID-19 precautions taken by hospitals around the United States via twitter data. In this project a total of 46 Illinois hospitals both rural and urban were taken into consideration. The total tweet dataset was about 11,000 dated from January 2020 to September 2020. The tools and techniques used were Twint for twitter data scraping, CorEx for topic modelling to find coherent meaningful topics measured across a corpus of text and finally an in depth exploratory analysis was performed to find the significant correlation between topics.A tweeting pattern that has made a considerable impact in generating awareness during COVID-19 by the hospitals is analysed.

## Table of Contents

## Introduction

2020 – a year where the unprecedented pandemic ‘COVID-19’ hit us hard killing millions around the world. There has been a worldwide crisis the likes of which has never been seen in the past hundred years. Doctors and nurses have now been busier in the trenches than ever before, fighting against the new invisible enemy. Cities are locked down and the common folk have been besieged in their own homes for a year now to prevent the spread of the virus. But at a time when everyone needs better information, there is a lack of reliable sources to guide decisions and actions of massive significance and to monitor their impact.

This project thus focuses on an analysis to identify what hospitals in the Illinois state have been tweeting about and to understand the kind of tweets and practices that have created an impact on people. This is done based on data streamed from Twitter from January 2020 until the end of September 2020. Using topic modelling, the topics around which the discussions can be classified, were identified and analyzed. Contrasting tweet practices were seen for every hospital, some being more popular than the others. The traction of a tweet depends on a variety of factors like the number of followers, the geographical location of the hospital (rural or urban), the age of the hospital twitter account, the day/time the tweet was created, etc. It also depends on the type of topic that the tweet belongs to.

Although hospitals tend to have preconceived notions about what they should be tweeting about it is vital to know what topics can create the most impact and develop a strategy to leverage them. How does one statistically define the best variables that have a direct impact? Should hospitals still follow the same tweet practices in the future if the pandemic ruffles across the globe unabated? How can hospitals tell if they are doing ‘more good’ than harm? The project focuses on answering some of these questions with regard to the best hospital tweet practices by mathematical simulations and regressions and concludes on optimal tweeting approaches for the hospitals, thus facilitating efficient communication and impact of various aspects of COVID-19 on public.

## Project Goals

**Goal 1:** To identify tweet-practices followed by hospitals in Illinois by analyzing their tweets during the COVID-19 pandemic.

**Goal 2:** To identify the factors that impact the traction of a tweet (number of likes and retweets) and recommend best tweet-practices to be followed by hospitals.

## Project Flow

**Goal 1:**

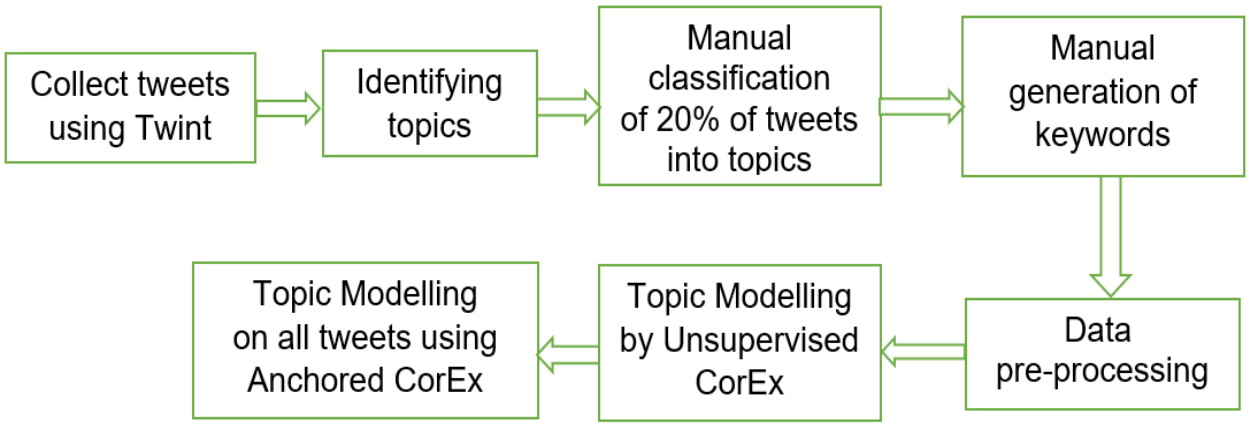


**Fig1. Project Flow for achieving Goal 1**

Steps illustrating Goal 1:

1. Collect tweets using twint- The tweets were collected using a twitter scraping tool in python, Twint. The total number of tweets collected were approximately 11000.
2. Identifying topics- 5% of the tweets (approx. 500 tweets) were selected in random order and the following topics were identified:
    a. Hospital Promotion- Hospital advertisements as well as general information about hospital timings and appointments
    b. Health Education/Awareness- General information regarding health and health care tips
    c. Event- Seminars, webinars and virtual events
    d. Award/Achievement/Appreciation- Awards and positive reviews received by the

hospital or hospital staff

e. Job Posting/Hiring- Open job positions
f. COVID specific event- COVID-19 specific seminars, webinars and virtual events
g. General COVID awareness- General COVID-19 specific information as well as health care tips from external sources such as CDC/WHO
h. Hospital specific COVID education- Hospital specific COVID-19 information as well as health care tips from hospital staff
i. COVID Policy Changes/hospital changes- New policies and changes in hospital rules and regulations introduced due to COVID-19
j. Miscellaneous- tweets that are not concerned with the health industry

3. Manual classification of 20% of tweets into topics- Once the topics were identified, additional 15% of the tweets were manually classified into the topics mentioned above. Thus, a total of 20% of the tweets were classified into topics.
4. Manual generation of keywords- Following the classification of the tweets into topics, the keywords for each topic was generated using word cloud. These keywords were mutually exclusive.
5. Data Pre-processing- The data was then pre-processed to remove emojis, urls, user mentions, punctuations and stop words. The tweets were also converted into lower case and lemmatization was performed on the data.
6. Topic Modelling by Unsupervised CorEx- The next step was to perform topic modelling on the data using Unsupervised CorEx, which is used to generate a set of keywords for user defined number of topics.
7. Topic Modelling on all tweets using Semi-Supervised CorEx- The keywords that were generated using word cloud were then supplied to Semi-Supervised CorEx model, which then classified the entire dataset into topics.

**Goal 2:**

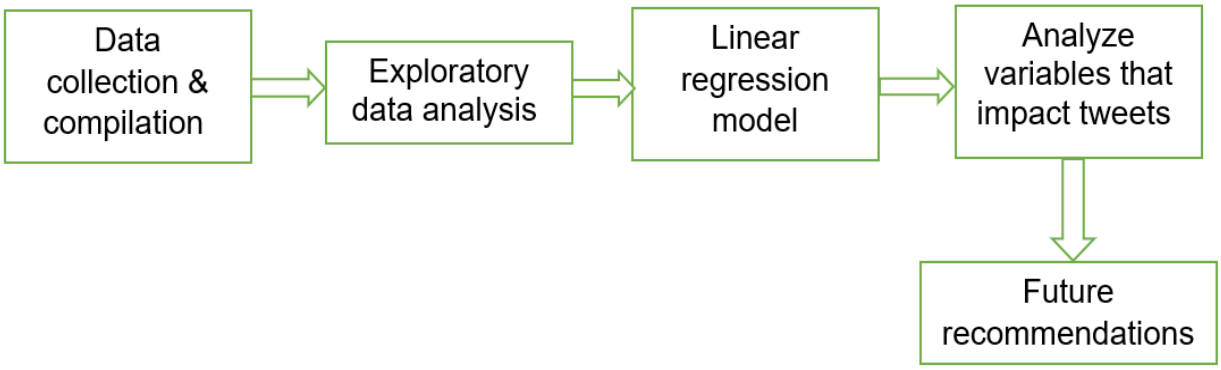


**Fig2. Project Flow for achieving Goal 2**

Steps illustrating Goal 2:

1. Data Collection and Compilation- This step involved the collection of additional data such as number of likes and retweets received by a tweet, as well as the tweet and organizational characteristics that impact the number of likes and retweets.
    a. Tweet Characteristics- Topic, Time of Day, Day, Media
    b. Organizational characteristics- Geographical Area, Number of Followers on Twitter, Days since account creation
2. Exploratory Data Analysis (EDA)- In this step, there are various methods used to perform some initial investigations on the data and discover certain patterns in it. The methods used are:
    a. Hypothesis Testing - Anova, t-test and Pearson Correlation
    b. Feature selection- Stepwise regression, Variance Inflation Factor (VIF), Correlation Matrix
3. Linear Regression Model- This is used to model the relationship between the number of likes, retweets received by the tweet and the factors that impact them.

## COVID Topic Analysis

Top subtopics hospitals tweeted about (in each COVID-related topic):

1. Hospital specific COVID education:
    a. Supporting front line workers
    b. How to prevent spread of COVID-19
    c. Encouraging people to get flu shot
    d. COVID-19 vaccine
2. COVID Specific event:
    a. COVID-19 drive-through testing events
    b. Events on vaccine clinical trials
    c. Online events on safely returning to school and work
    d. Online events on precautions and best practices to be taken to prevent spread of COVID-19
    e. Online events on recovering from COVID-19
3. COVID Policy Changes:
    a. COVID hotline
    b. Visitor policy
    c. Patient or visitor screening
4. General COVID Awareness:
    a. CDC guidelines and recommendations
    b. Traveling during COVID-19
    c. How cancer patients can take extra precautions to protect themselves from COVID-19

## Tools and Techniques

### 1. Twint

Scraping Twitter with Twint, the official Twitter API is effective in certain respects, but the amount and speed of data collection is fairly restrictive. Twitter's Basic Search API only goes back seven days, while social scientists are also interested in data from months or years ago. The cost of buying historical Twitter data is always out of the control of the average social scientist, and only limited data can be scraped using Twitter API.

Hence twitter scraping was implemented using Twint. The Twint Python library takes advantage of this so that the Twitter information can be collected without using the API. Twint was used for scraping the data for 58 hospitals, out of which only 46 hospitals had tweets dated from January 1 to September 30, 2020.

### 2. CorEx

The principle of Cor-relation Ex-planation is a way to build rich representations that are informative about relationships in data, the two-thirds of the tweets collected by twint were manually segmented into topics and CorEx was then used for finding coherent meaningful topics measured across a corpus of text.

The major dependencies of CorEx is it requires numpy and scipy, segmenting and topic modelling was done using two methods Unsupervised and Semi-Supervised learning.

Unsupervised topic modelling using CorEx is the way to generate a set of top ten keywords for corresponding topics.

Semi-Supervised topic Modelling using CorEx uses manually generated keywords and the anchor key words topic which was generated in unsupervised learning which were mutually exclusive for every topic.

## 3. Exploratory Data Analysis

### a. Distribution of dependent variables (Likes and Retweets)

WITH OUTLIERS

| | Likes | Retweets |
|---|---|---|
| count | 11061.00 | 11061.00 |
| mean | 4.29 | 1.33 |
| std | 16.76 | 5.20 |
| min | 0.00 | 0.00 |
| 25% | 0.00 | 0.00 |
| 50% | 1.00 | 0.00 |
| 75% | 4.00 | 1.00 |
| max | 857.00 | 341.00 |

OUTLIERS REMOVED

| | Likes | Retweets |
|---|---|---|
| count | 10938.00 | 10938.00 |
| mean | 3.21 | 1.01 |
| std | 5.86 | 1.86 |
| min | 0.00 | 0.00 |
| 25% | 0.00 | 0.00 |
| 50% | 1.00 | 0.00 |
| 75% | 3.00 | 1.00 |
| max | 54.00 | 16.00 |

**Fig3. Distribution of Likes and Retweets (With and Without Outliers)**

The above picture provides the descriptive statistics for the likes and retweets with the initial data and once outliers have been removed. With the removal of outliers the mean likes drops from 4.29 to 3.21 and mean retweets reduces from 1.33 to 1.01.

**b. Hypothesis testing:** Statistical techniques conducted to find significant variables in the dataset. ANOVA, t-Test and Pearson correlation has been used to perform hypothesis testing.

**(i) ANOVA:**

Analysis of Variance is used to compare means of more than two groups, working of ANOVA is as following:

- Checks sample size and makes sure that equal number of observations are there in each group
- Calculates the mean square of each group (MS)
- Calculates the Mean Square Error (MSE)

The p value obtained from ANOVA analysis indicates p value< 0.05 to be significant variables.

Anova was used to check the statistical significance of Time of Day and Topic on the number of likes and retweets.
According to Anova, Time of Day has no statistical significance on the number of likes and retweets, whereas Topic has a statistical significance on the number of likes and retweets.

**(ii) t-Test:**

The independent T-test is a test used to analyse for a statistically significant difference in the means between 2 groups.
The larger the t score, the more difference there is between groups. The smaller the t score, the more similarity there is between groups.

t-Test was used to check the statistical significance of Day, Geographical area and Media on the number of likes and retweets.
According to t-Test, Day and Geographical area has no statistical significance, on the number of likes and retweets, whereas Media has a statistical significance on the number of likes and retweets.

**(iii) Pearson's Correlation:**

The correlation coefficient for Pearson is the test statistics that calculate the statistical relationship between two continuous variables, or association.
Since it is based on the covariance principle, it is regarded as the best method of calculating the correlation between variables of interest.

Pearson's Correlation was used to check the statistical significance of Days since account creation and Number of Followers on the number of likes and retweets.

According to Pearson's Correlation, both Days since account creation and Number of Followers have a statistical significance on the number of likes and retweets.

c. **Feature selection:** A process which automatically selects features most relevant to the prediction variable or output. Having irrelevant features in the prediction model reduces accuracy and makes the model learn on irrelevant features. Stepwise regression was used for feature selection in the project.

**(i) Stepwise regression:**

The backward elimination model has been used under the stepwise regression method. The backward elimination model begins with all the variables in the model. At every step, the variable that is the least significant is being removed. This process is continued until there are no more nonsignificant variables. The significance level at which variables can be removed from the model is set by the user.

The statistically significant variables found from the stepwise regression for number of likes are Number of Followers, Days since account creation, Award/Achievement/Appreciation, COVID Policy Changes/hospital changes, COVID specific event, Event, General COVID awareness, Health Education/Awareness, Hospital Promotion, Hospital specific COVID education, Miscellaneous, No Media, Rural, Urban, Afternoon, Evening, Morning, Night, Weekday, Weekend.

The statistically significant variables found from the stepwise regression for number of retweets are Number of Followers, Days since account creation, Award/Achievement/Appreciation, COVID Policy Changes/hospital changes, COVID specific event, Event, General COVID awareness, Health Education/Awareness, Hospital Promotion, Hospital specific COVID education, Miscellaneous, No Media, Urban, Afternoon, Evening, Morning, Night,Weekday, Weekend.

From stepwise regression, it was found that some of the independent variables had high intercorrelations (multicollinearity). To address the multicollinearity issue, VIF(Variance Inflation Factor) scores were calculated by eliminating some of the collinear variables or variables with high VIF scores.

**(ii) Variance Inflation Factor (VIF):**

The VIF in a set of multiple regression variables is a measure of the amount of multicollinearity**.**

The statistically significant variables found from the VIF for number of likes and retweets are Number of Followers, Days since account creation, COVID Policy Changes/hospital changes, COVID specific event, Event, General COVID awareness, Health Education/Awareness, Hospital specific COVID education, Job Posting Hiring, Miscellaneous, No Media, Urban, Evening, Morning, Night, Weekend.

To see variables with high collinearity, a correlation matrix was used as a visual approach.

**(iii) Correlation matrix:**

A table displaying correlation coefficients between variables is a correlation matrix. The association between two variables is presented by each cell in the table. To summarize results, as an input into a more advanced analysis, and as a diagnosis for advanced analyses, a correlation matrix is used.

**Final Selection of Features:** Based on the above Feature selection analysis, the statistically significant variables found for number of likes are Number of Followers, Days since account creation, COVID Policy Changes/hospital changes, COVID specific event, Event, General COVID awareness, Health Education/Awareness, Hospital specific COVID education, Miscellaneous, Contains Media, Urban, Evening, Morning, Night, Weekend.

Based on the above Feature selection analysis, the statistically significant variables found for number of retweets are Number of Followers, COVID Policy Changes/hospital changes, Hospital specific COVID education, Job Posting Hiring, Miscellaneous, No Media, Urban.

4. **Linear Regression:**

A linear regression model aims at finding a correlation between one or more characteristics (independent variables) and a continuous target variable (dependent variable).

## Results and Analysis

### 1. Topic Modelling Output

The Semi-supervised or Anchored CorEx classifies the entire dataset into topics based on the anchor words provided and generates the following set of keywords for each topic:

```
0: extensive onsite,position,certification,join team,looking hire,food service,certified
1: care wondering,healthcare caring,provide,medical center,robotic surgery,immediate care,carle_docs
2: blood pressure,heart health,screening,colorectal cancer,heart attack,cancer awareness,awareness
3: thank staff,appreciation,thank nurse,award,care patient,help ensure,thank
4: discussion,megan wright,completed fall,meiner,flu shot,time register,crate
5: symptomatic,non symptomatic,clinic,interested tested,care connect,care eye,zvi
6: room main,rooney md,healthy relationship,wearing mask,social distancing,social distance,teen facing
7: dr,expert,chief medical,answered,officer dr,askdrmarkman,disease expert
8: screening,visitor policy,safety patient,visitor staff,policy,new visitor,health safety
9: census,relation team,football,experienced patient,apologize dissatisfaction,patient relation,happynewyear
```

**Fig4. Semi-supervised Output**

### 2. Distribution of Topics

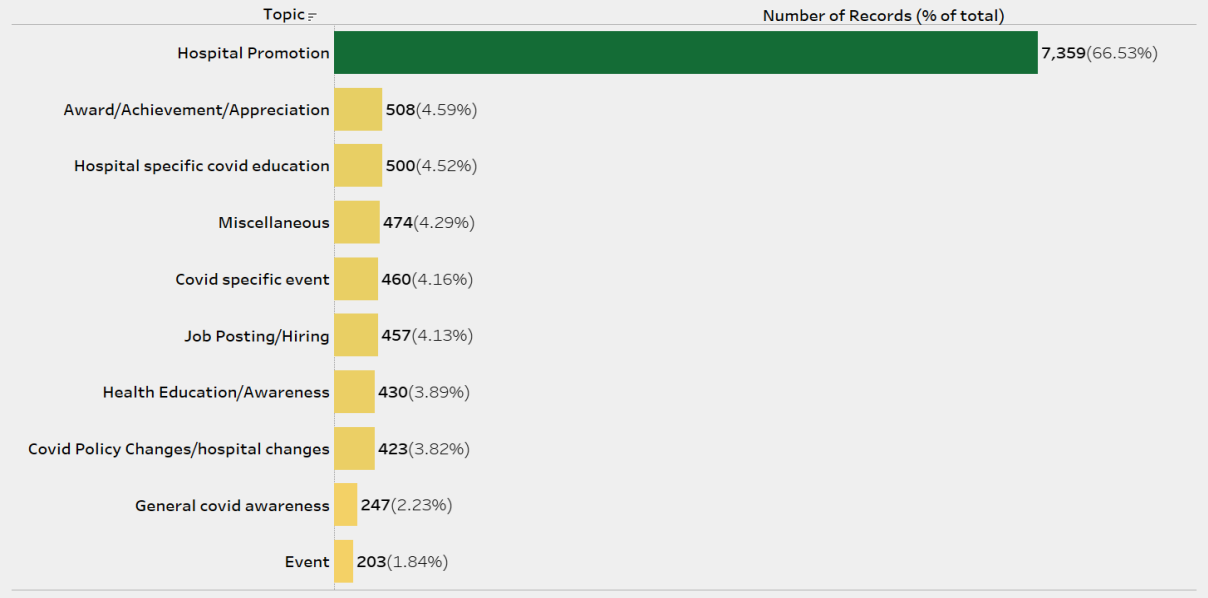


**Fig5. Distribution of Topics**

a. Two-thirds of tweets by Illinois Hospitals belong to the Hospital Promotion followed by Award/Achievement/Appreciation
b. Event is the least tweeted Topic
c. COVID related tweets are equally distributed between Hospital specific COVID

education and COVID specific event

d. General COVID awareness is the least tweeted COVID Topic

## 3. COVID Tweet Visualizations

The COVID related Tweet Visualizations are shown below:

| Hospital | Topic: Covid Policy Changes/hospital changes | Covid specific event | General covid awareness | Hospital specific covid education |
|---|---|---|---|---|
| Cancer Treatment Centers of America (CTCA) | 16 | 67 | 5 | 71 |
| Franciscan Health | 31 | 24 | 39 | 39 |
| Riverside Healthcare | 19 | 26 | 61 | 23 |
| Rush University System for Health | 26 | 38 | 13 | 38 |
| Advocate Health Care | 31 | 32 | 15 | 30 |
| Sinai Health System | 21 | 19 | 6 | 17 |
| Carle | 25 | 14 | 4 | 10 |
| OSF HealthCare | 20 | 8 | 4 | 14 |
| SSM Health | 8 | 17 | N/A | 11 |
| Northwestern Medicine | 6 | 4 | 9 | 10 |

Number of Tweets: 4 – 71

**Fig6. Distribution of the COVID related topics**

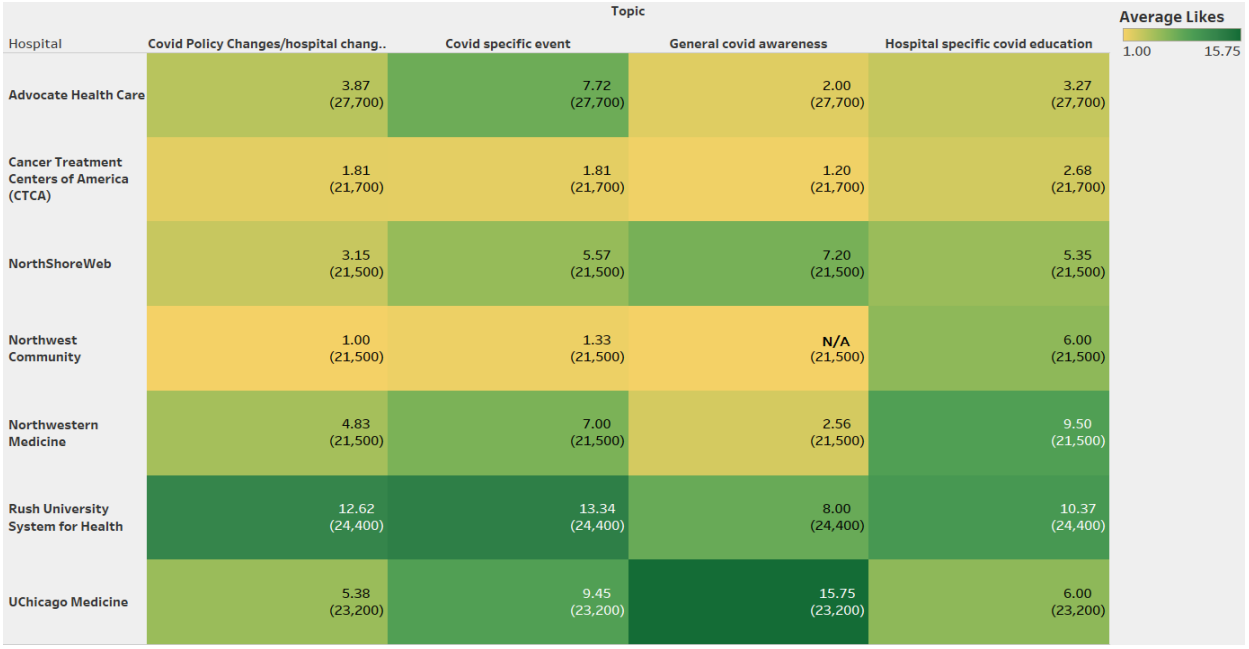


**Fig7. Average Likes for COVID Tweets**

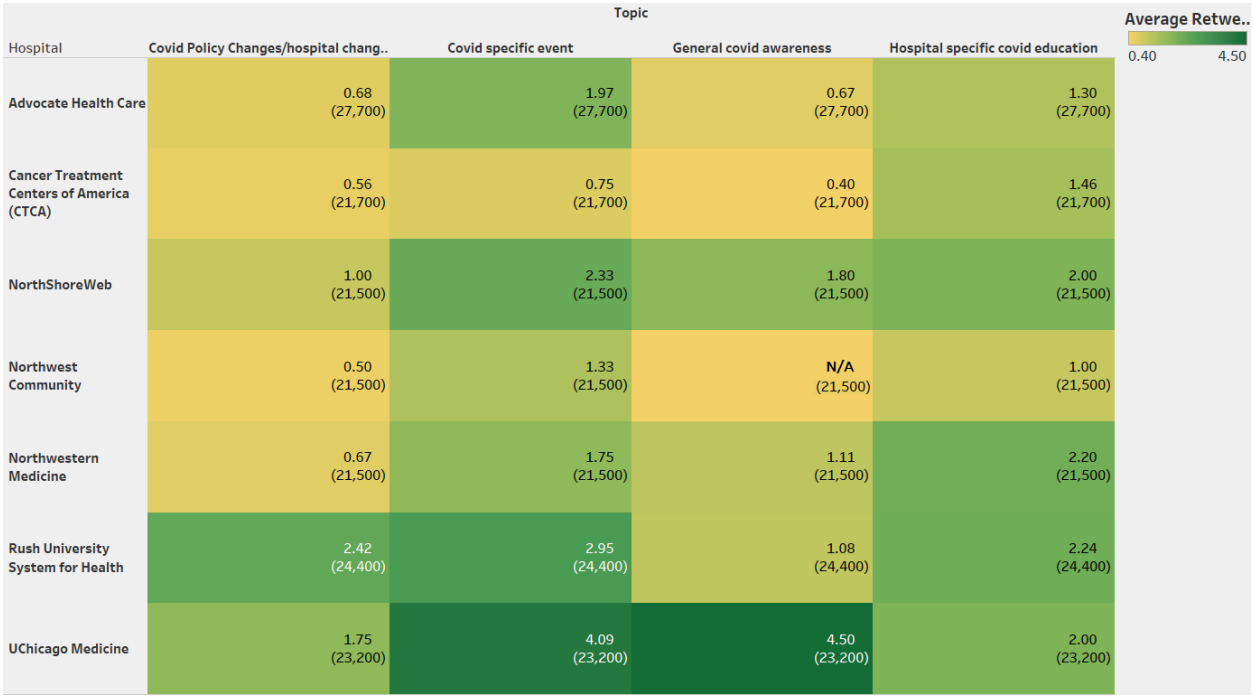


**Fig8. Average Retweets for COVID Tweets**

Note for Fig7 and Fig8: The values in brackets are number of followers for each hospital > 20k

## 4. Linear Regression Output

a. Dependent Variable: Likes

Results: Ordinary least squares

| | | | |
|---|---|---|---|
| Model: | OLS | Adj. R-squared (uncentered): | 0.114 |
| Dependent Variable: | Likes | AIC: | 84659.1195 |
| Date: | 2020-12-10 19:07 | BIC: | 84767.2055 |
| No. Observations: | 9954 | Log-Likelihood: | -42315. |
| Df Model: | 15 | F-statistic: | 85.99 |
| Df Residuals: | 9939 | Prob (F-statistic): | 2.04e-249 |
| R-squared (uncentered): | 0.115 | Scale: | 288.78 |

| | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Number of Followers | 8.7649 | 0.5347 | 16.3932 | 0.0000 | 7.7168 | 9.8130 |
| Days since account creation | -2.7135 | 0.4924 | -5.5104 | 0.0000 | -3.6787 | -1.7482 |
| Covid Policy Changes/hospital changes | 0.9879 | 0.8982 | 1.1000 | 0.2714 | -0.7726 | 2.7485 |
| Covid specific event | 2.8129 | 0.8623 | 3.2621 | 0.0011 | 1.1226 | 4.5032 |
| Event | 3.3596 | 1.2688 | 2.6478 | 0.0081 | 0.8725 | 5.8468 |
| General covid awareness | 0.4133 | 1.1410 | 0.3623 | 0.7171 | -1.8232 | 2.6498 |
| Health Education/Awareness | 4.5679 | 0.8806 | 5.1872 | 0.0000 | 2.8417 | 6.2941 |
| Hospital specific covid education | -1.3975 | 0.8294 | -1.6849 | 0.0920 | -3.0233 | 0.2283 |
| Miscellaneous | -0.4927 | 0.8402 | -0.5865 | 0.5576 | -2.1396 | 1.1541 |
| Contains Media | 3.7470 | 0.3517 | 10.6536 | 0.0000 | 3.0576 | 4.4364 |
| Urban | 4.1233 | 0.3985 | 10.3474 | 0.0000 | 3.3422 | 4.9044 |
| EVENING | -0.2947 | 0.3502 | -0.8415 | 0.4001 | -0.9812 | 0.3918 |
| MORNING | 0.5213 | 1.6607 | 0.3139 | 0.7536 | -2.7341 | 3.7766 |
| NIGHT | -0.2774 | 0.6782 | -0.4091 | 0.6825 | -1.6069 | 1.0520 |
| Weekend | 0.5464 | 0.4879 | 1.1200 | 0.2627 | -0.4099 | 1.5027 |

| | | | |
|---|---|---|---|
| Omnibus: | 25096.409 | Durbin-Watson: | 2.006 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 574114430.425 |
| Skew: | 27.679 | Prob(JB): | 0.000 |
| Kurtosis: | 1178.235 | Condition No.: | 12 |

**Fig9. Linear Regression for Likes**

(i) Since the Adjusted R-squared value has increased from 0.059 (using all variables) to 0.114 (chosen variables), thus variables that were chosen have significantly improved the model.

(ii) AIC has decreased from 93087.734 to 84659.1195 from that of the previous model indicating that the regression model has improved prediction accuracy.

(iii) The variables that are statistically significant at 95% confidence interval are Number of followers, Days since account creation, COVID specific event, Event, Health Education/Awareness,Contains Media, Urban.

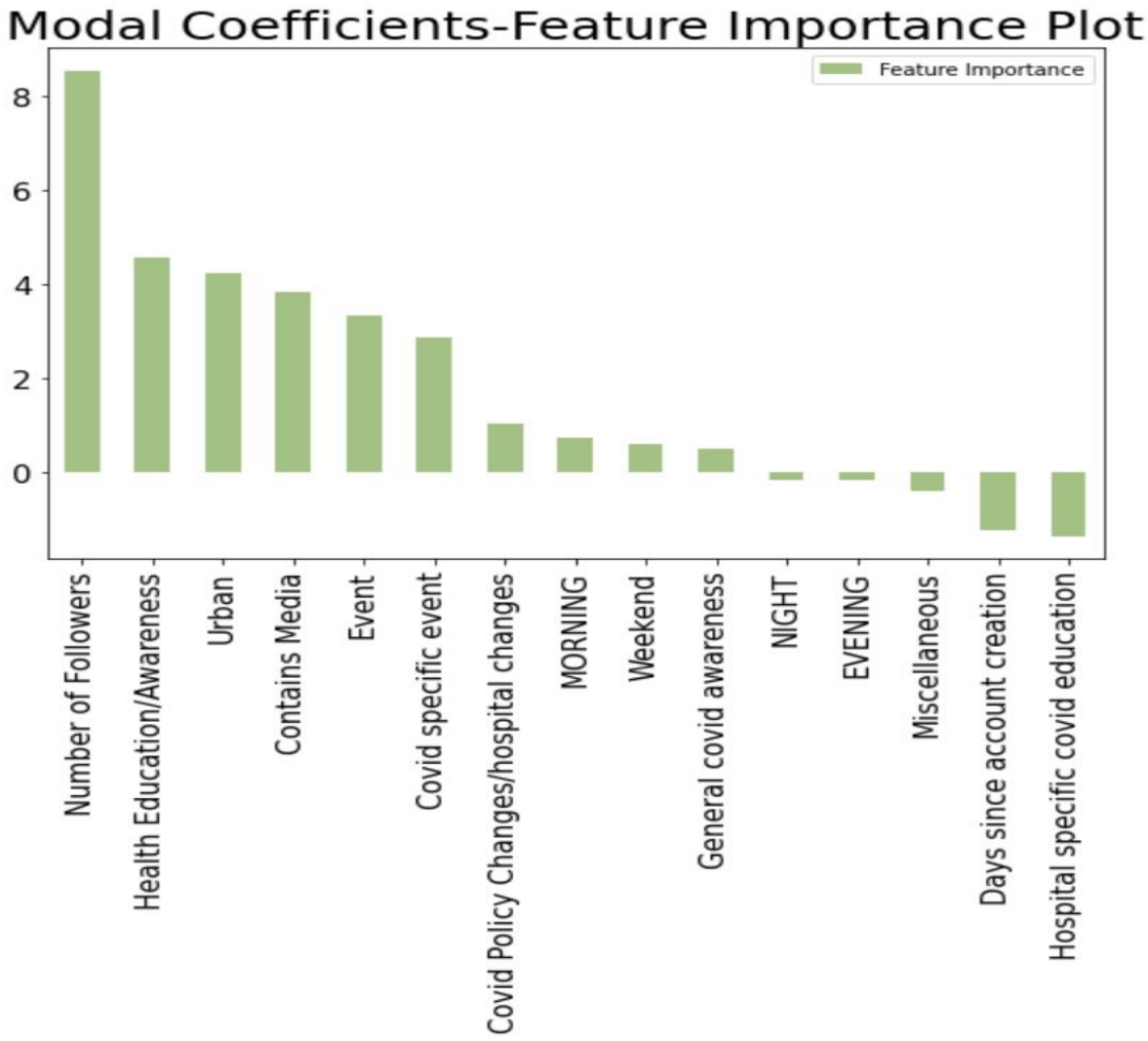


**Fig10. Feature Importance for Likes**

Based on the Linear Regression model for likes, the most important variables found for number of likes are Number of Followers, Health Education/Awareness, Urban, Contains Media, Event, COVID Specific Event, COVID Policy Changes/hospital changes, Morning, Weekend, General COVID Awareness.

b. Dependent Variable: Retweets

```
                          Results: Ordinary least squares
================================================================================
Model:                      OLS              Adj. R-squared (uncentered):  0.093
Dependent Variable:         Retweets         AIC:                          61490.6085
Date:                       2020-12-10 19:07 BIC:                          61541.0486
No. Observations:           9954             Log-Likelihood:               -30738.
Df Model:                   7                F-statistic:                  147.4
Df Residuals:               9947             Prob (F-statistic):           9.14e-208
R-squared (uncentered):     0.094            Scale:                        28.188
--------------------------------------------------------------------------------
                                    Coef.  Std.Err.    t    P>|t|   [0.025  0.975]
--------------------------------------------------------------------------------
Number of Followers                 2.4930   0.1335 18.6742 0.0000  2.2313  2.7547
Covid Policy Changes/hospital changes  0.2151   0.2759  0.7795 0.4357 -0.3258  0.7559
Hospital specific covid education   0.0250   0.2545  0.0984 0.9216 -0.4738  0.5239
Job Posting/Hiring                 -0.2526   0.2649 -0.9534 0.3404 -0.7719  0.2667
Miscellaneous                       0.0696   0.2592  0.2687 0.7882 -0.4385  0.5778
No Media                           -0.5795   0.0967 -5.9946 0.0000 -0.7690 -0.3900
Urban                               1.6398   0.1196 13.7158 0.0000  1.4054  1.8741
--------------------------------------------------------------------------------
Omnibus:                27489.081          Durbin-Watson:             2.006
Prob(Omnibus):          0.000              Jarque-Bera (JB):          1511036274.183
Skew:                   34.817             Prob(JB):                  0.000
Kurtosis:               1910.458           Condition No.:             4
================================================================================
```

**Fig11. Linear Regression for Retweets**

(i) Since the Adjusted R-squared value has increased from 0.059 (using all variables) to 0.093 (chosen variables), thus variables that were chosen have significantly improved the model.

(ii) AIC has decreased from 67376.358 to 61490.6085 from that of the previous which indicates that the model has improved prediction accuracy.

(iii) The variables that are statistically significant at 95% confidence interval are Number of followers, No media, Urban.

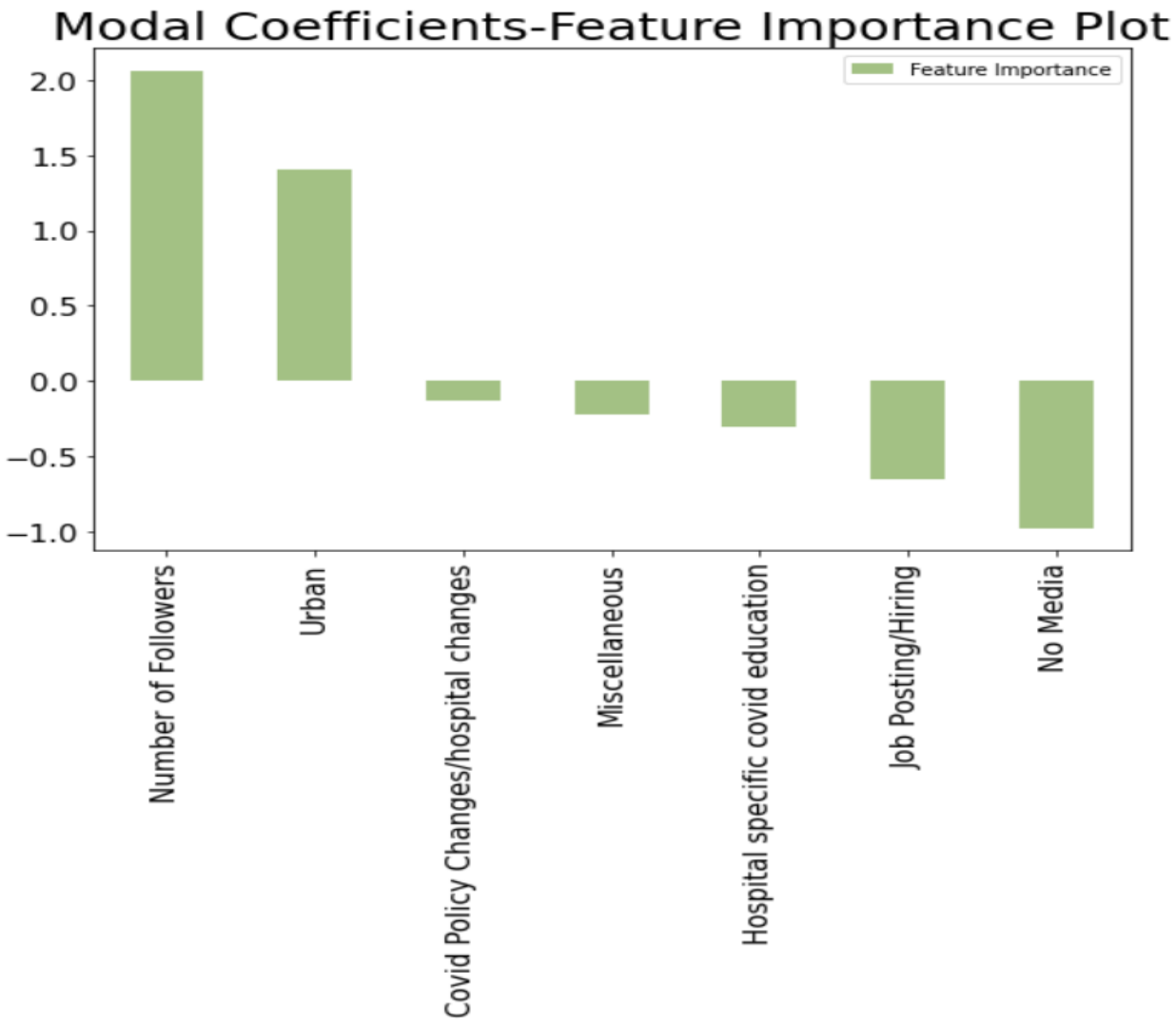


**Fig12. Feature Importance Plot for Retweets**

Based on the Linear Regression model for retweets, the most important variables found for the number of retweets are Number of Followers and Urban.

## Recommendations

1. Hospitals should consider tweeting on topics like Health Education/Awareness, COVID Specific Event and Event, since they seem to receive higher user engagement
2. Tweets including media such as image or video seem to have a higher popularity
3. Although obvious, increasing the number of twitter followers for a hospital twitter account plays a pivotal role in widening their reach
4. Tweeting about Hospital specific COVID education seems to generate less traction as compared to other COVID related tweets
5. Miscellaneous and Job posting/hiring tweets should be avoided as they have the least impact
6. Reducing direct promotional tweets and instead focusing on tweeting about Health education/Awareness and Event

## Conclusion

With the current pandemic luring around with an indefinite end, it is imperative that proper measures be taken by the hospitals and most importantly, that the hospitals follow mostly those social media (tweeting) practices which have a higher user engagement. Following those practices would not just help people get the right kind of information from the right source but would also help the hospital gain traction on their tweets, while building their brand name at the same time. From the analysis of this project, hospitals should engage in tweeting more about health education and awareness and should have more events. In times of the pandemic, COVID specific events have proven to be quite popular. Hospitals could also focus on tweeting only about hospital related tweets because miscellaneous tweets do not gain much popularity. This project has, therefore, focused on finding impactful tweet practices and provided recommendations to help the medical institutions understand their reach better.

**References**

Sharma, Himanshu. *Complete Tutorial On Twint: Twitter Scraping Without Twitter's API* https://analyticsindiamag.com/complete-tutorial-on-twint-twitter-scraping-without-twitters-api/

Zacharias, Cody (2020, April 29). *TWINT - Twitter Intelligence Tool*. https://pypi.org/project/twint/

Gallagher, Ryan J. (2018, July 21). *Anchored CorEx: Topic Modeling with Minimal Domain Knowledge*.https://github.com/gregversteeg/corex_topic/blob/master/corextopic/example/corex_topic_example.ipynb

Ver Steeg, Greg (2017, June 2). *Correlation Explanation (corEx)* https://github.com/gregversteeg/CorEx

Amulya, Aankul(2017, Aug 30) *T-test using Python and Numpy* https://towardsdatascience.com/inferential-statistics-series-t-test-using-numpy-2718f8f9bf2f

Glen, Stephanie. *"ANOVA Test: Definition, Types, Examples*" From StatisticsHowTo.com: https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/

NCSS. *Stepwise Regression*. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf

Hascelik, Sinan Talha (2019, April 23). *Automated Stepwise Backward and Forward Selection*. https://github.com/talhahascelik/python_stepwiseSelection

Beck, Marcus W (2013). *Collinearity and stepwise VIF selection*. https://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/