Milestone 1: Report

Coronavirus Tweet Categorization and Exploration

Group 2: - Lisa Peterson, Gabriel Cruz, Shreeya Kotasthane, Md Main Uddin Rony

INTRODUCTION

Social media has become one of the primary platforms for the expression of people's opinions. The intention of this project is to gather tweets that concern the recent coronavirus outbreak and gain insights from public opinions to obtain the answers to questions regarding the validity of the information that presents itself online. More specifically, the concern is that such large platforms can give way to the proliferation of falsehoods and conspiracy theories. Social media platforms like twitter make it easy for regular people to spread information, whether willingly or not, to a large population. Racist or biased ideas may also be proliferated too. Misinformation about entire countries or populations could be spread by these means as well. It is important to understand then how a health concern such as the coronavirus is presented online, and how the information that users are exposed to guides the information seeking behaviours on these social media platforms such as Twitter.

A. RESEARCH QUESTIONS

1. **Can we predict the number of the retweet or favorite count based on the tweet text and other metadata?**
    - This question brings awareness to what sort of tweets or opinions are common during a health pandemic. From a societal perspective, this question helps us to shed light on the real concerns of the public. By predicting these concerns, the government and the concerned authorities can save a certain amount of feelings and disconcert and stress amongst the public by addressing these issues beforehand.
    - Popularity prediction of tweets has been used in many fields. This question helps us to use this machine learning application to help solve concerns about the coronavirus.
2. **Can we predict if the tweet is fake/rumor/hoax using a classification method?**
    - A lot of chaos is caused during health pandemics. People are generally concerned about questions like whether or not it will spread in the area they live in and how they will be affected and they want to know if there are any measures they can take to ensure their safety. Because of these feelings of unrest and confusion amongst people, they are vulnerable to misunderstandings and rumors. To prevent anyone from taking misadvantage of this situation, to promote their propaganda or abuse it for financial advantages, classification of the tweets into true and hoax content becomes necessary.
    - Fake news detection is a very trending application of machine learning. A lot of research has been done with regards to this topic. Machine Learning and Deep Learning models have both been used in this field. Confirming the veracity of

tweets supports this research, as people often obtain news and information from social media.

3. **Do the fear causing tweets get more attention during an epidemic like Coronavirus than fear relieving tweets?**
    - Epidemics can cause people to fear and be more aware of the information that they consume as well as how they consume it. It is important to note that while the goal isn't to suppress fear inside of a population, it is worth the effort to gather information as to what exactly is causing populations to fear an epidemic. If the main source of fear is content that they consume online, then it would be helpful for government organizations to understand what it is specifically that is causing fear within the population in order to work against that concern.
    - This can be monitored using sentiment analysis and this relates to previous research done on monitoring health concerns.
4. **Can we predict if Twitter users are asking for any information about the Coronavirus? If so, what type of information are users looking for?**
    - Information seeking behaviours are very important to understand during global events since it is well known that there will be a subset of individuals that will not entirely attend to their concerns via legitimate channels. By identifying these tweets, the concerned authorities can provide clarification of doubts and maybe even provide assurance by way of giving accurate information.

B. STATE OF THE ART

- ***Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak***

The 2009 outbreak of the H1N1 influenza was the first pandemic that transpired during the Web 2.0 era. It was the first time social media was used as the primary source of information by the public. Not only did it allow for dissemination of factual resources but became a source that allowed the public to engage their own fears and experiences but to spread fallacies as well. Monitoring of such activity through outlets like Twitter has provided an immediate medium that goes far beyond the traditional standards of print, television, and surveys. Data mining of social media activity, often referred to as “infoveillance”, can provide the medical community with real-time analysis of public perception and allow for strategizing of communications.

* * *

Chew C, Eysenbach G (2010) Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE 5(11): e14118. doi:10.1371/journal.pone.0014118.

- ***You Are What You Tweet: Analyzing Twitter for Public Health***

Analysis of social media has developed beyond the tracking of a single ailment. Infoveillance of public health has broadened to analysis of multiple ailments geographically, behavioral risks, and symptoms with treatments. Paul and Dredze expanded the ATAM (Ailment Topic Aspect Model) to include social media data and expand public health informatics. Monitoring data of millions of tweets allows for “significant impact on public health, impacts medical resource allocation,

health policy and education." With Twitter users sharing comments of how they're feeling, medicines and therapies being taken, along with positive or negative outcomes, a wealth of information is made available to the health and science community. Information analysis derived from Twitter provides an additional benefit of being less expensive and time consuming than traditional methods, but cons are the demographics of Twitter users are predominantly younger and may not be as commonly used in other countries.

* * *

Paul M, Dredze M (2011) You Are What You Tweet: Analyzing Twitter for Public Health. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (pp. 265–272). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880/3264.

- ***Effects of misleading media coverage on public health crisis: a case of the 2019 novel coronavirus outbreak in China***

  Media coverage has a large impact on primarily three areas. First, misleading media coverage can impact the spread of the outbreak by misleading the public into practices that can aid the spread of the outbreak. Practices like misnaming and misinforming the public about how serious an outbreak is could cause the public to take a more relaxed stance to a serious situation. Secondly, misleading information could exhasterbate racial discrimination and cause people to connect an outbreak with a race and not with the health issues surrounndng it. This sort of discriminiation could cause mental health disorders to show themselves as well in this discriminated populations. Lastly, misinformation can negatively impact the public's view of different destinations as tourist areas. This could in turn affect the local economy of areas where the economy might already be fragile due to the illnesses. China's NCP outbreak can very similarly follow these three patterns if misinformation has the chance to spread across the different channels.

  * * *

  Jun Wen, Joshua Aston, Xinyi Liu & Tianyu Ying (2020) Effects of misleading media coverage on public health crisis: a case of the 2019 novel coronavirus outbreak in China, Anatolia, DOI: 10.1080/13032917.2020.1730621

- ***Forecasting the Future: Leveraging RNN based Feature Concatenation for Tweet Outbreak Prediction***

  Different methods exist for classifying tweets based on the features that they contain. A proposed method for determining whether a post will be popular is to use a RNN (recurrent neural network) model. The RNN is set up in a way where the feature extractor normalizes data from the social aspects of the tweet such as who follows the tweeters account and how many tweets that have sent out in total. This sort of RNN is able to outperform other models by a significant amount and is there for a method to take a look at for being able to determine if a tweet will garner major traction online or not.

  * * *

  Saswata Roy, Brijendra Suman, Joydeep Chandra, Sourav Dandapat, Saswata Roy (2020) Forecasting the Future: Leveraging RNN based Feature Concatenation for Tweet Outbreak Prediction (pp. 219–223). Retrieved from https://dl.acm.org/doi/abs/10.1145/3371158.3371190

- ***Monitoring Public Health Concerns Using Twitter Sentiment Classifications***
  Twitter can serve as an important data source into the current opinion of the public. This paper focuses on the Degree of Concern amongst twitter users by sentiment analysis. In order to achieve this goal, we develop a novel two-step sentiment classification workflow to automatically identify personal tweets and negative tweets. This paper talks about an Epidemic Sentiment Monitoring System (ESMOS) that provides tools for visualizing Twitter users' concern towards different diseases. There is a visual concern map proposed in the ESMOS to help perform analysis regarding peak concerns with regards to location and time and take preventative measures. The DOC measure is based on sentiment-based classifications. Different ML methods are compared in order to classify sentiments of Twitter users regarding diseases, first into personal and neutral tweets and then into negative from neutral personal tweets.

  * * *

  X. Ji, S. A. Chun and J. Geller, "Monitoring Public Health Concerns Using Twitter Sentiment Classifications," *2013 IEEE International Conference on Healthcare Informatics*, Philadelphia, PA, 2013, pp. 335-344. doi: 10.1109/ICHI.2013.47

C. DATASETS

At present, there is no available dataset that contains tweets related to Coronavirus. So, we took the initiative to collect the tweets through Twitter Streaming API[1]. Using a python package named tweepy[2], we wrote a python script to collect the streaming tweets from February 6, 2020 to February 10, 2020. We ran the script with only one keyword which is *"#coronavirus"*. As tweet searching is case insensitive, we didn't have to consider the variations of these hashtags. To make sure that our collection only contains English tweets, we filtered out the tweets of other languages. We also filtered out the retweets by checking if the tweet text contains 'RT' in it or not. We parsed 17 fields from the API response which we think might be interesting. They are tweet text, tweet creation time, tweet id, tweet URL, tweet source, user's name, screen name, user's location, user's verification information, user's follower count, quote status, quote count, reply count, retweet count, favorite count, hashtags, and user mentions. We used a relational database management system, SQLite[3], to store the streaming data. In total, we collected 1,75,251 tweets related to coronavirus.

Our exploratory data analysis showed 76,408 unique users who tweeted the tweets and among them 2,930 accounts are verified. Table 1 shows the top 5 accounts in terms of tweets count. One interesting observation is that none of the top 5 accounts is verified. Verified users posted 10,285 tweets with an average 3.5 tweets/user which is greater than normal average (2.29 tweets/user).

User Name	Tweet Count	Account Verified
bitcoinconnect	4578	No

[1] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data
[2] http://docs.tweepy.org/en/v3.5.0/api.html
[3] https://www.sqlite.org/index.html

EcoInternetDrGB	591	No
TPE_connect	498	No
anmedia6	430	No
vistabuzz	390	No

Table 1: Top 5 Twitter accounts (Ranked by Tweet Counts)

We also did some statistical analysis of users' follower count, retweet count and favorite count. Table 2 shows them in detail. Although the follower count analysis is done on the whole dataset, retweet and favorite count analysis is done on 18,000 tweets only. The reason behind this discrimination is tweets are collected through streaming API. So they are stored in our database as soon as they are available on Twitter with less number of retweets and favorite count (almost all of them were showing count 0). So, we ran another python script for accessing the Twitter API to get the new tweet's information of our stored tweets on February 24, 2020 (almost 2 weeks after first round collection). Tweet API has some restrictions on continuous API accessing[4], so the second round data collection wasn't fast enough to get all the tweet's information. We could get the information of 18K tweets only before writing this report.

Filed	Max	Min	Avg	25%	50%	75%
Follower	23659780	0	9603.63	74	391	1696
Retweet	4384	0	4.51	0	0	1
Favorite	8737	0	8.79	0	1	2

Table 2: Statistical details of Users' follower count, retweet count, and favorite count

We also investigated the top hashtags and user mentions for our collected tweets. Top hashtags can help us to download more data using them if needed. One interesting observation could be checking the tweets with the hashtags like "#fake", "#fakenews", "#falsenews", "#hoax", etc, because these tweets will be helpful for addressing our RQ3. We got 12, 205, 1, and 10 tweets for the above mentioned hashtags. Table 4 and 5 shows the top 5 hashtags and user mentions.

Hashtags	Count
#coronavirus	83545
#china	10160

[4] https://developer.twitter.com/en/docs/basics/rate-limiting

#wuhan	5666
#coronavirusoutbreak	3280
#virus	3170

Table 3: Top Hashtags

User Mentions	Count
@who	1678
@realdonaldtrump	1025
@youtube	781
@drtedros	471
@cdcgov	421

Table 4: Top User Mentions

We did some textual analysis on the text of the tweets. Before doing this we cleaned the text first. The cleaning process is described under the Data Cleaning Efforts section. We did the analysis on 18K tweets because after the first round of data collection we found that many tweet texts we got are truncated. So in our second round of data collection, we collected full text. The average length of the tweet is 18.90 words. We also took some interest to check the most occured words in the tweets. Figure 1 shows the word clouds built from the tweet texts. We can see the prevalence of the words like “China”, “virus”, “cruise”, “outbreak”, etc, from the clouds. During the curation of the tweets, the news of discovering coronavirus affected people in a cruise broke out. That might be the reason for “cruise” to being one of the influencing words in the cloud.

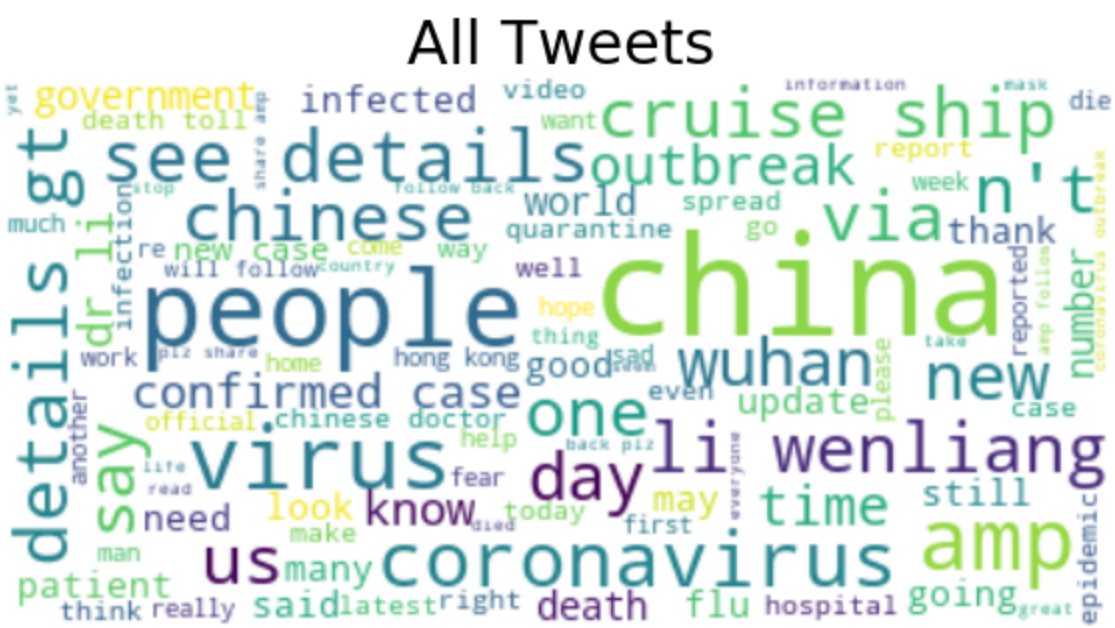

Figure 1: Word Cloud of the tweets

D. DATA CLEANING EFFORTS

For cleaning the tweets, we have taken two approaches. In the first approach, we used a python package named "tweet-preprocessor"[5] to clean the tweets initially. Then we used a manual approach to make sure tweets are cleaned actually. We tokenized the tweet text using python NLTK package[6], then removed the punctuation and emoji symbols. To make the word cloud meaningful, we also removed the stop words from the tokens. There were some unwanted non-ASCII characters which we removed using Regex.

E. OTHER SOFTWARE ENGINEERING EFFORTS

Twitter streaming API doesn't give the URL of the tweets as a response field. So we took a manual approach to build the URL from tweet id and user name. The common pattern of the url is: https://twitter.com/user_name/status/tweet_id. Also we may need to put some extra effort in future to merge the data collected in two rounds.

[5] https://pypi.org/project/tweet-preprocessor/
[6] https://www.nltk.org/