Zhihan Zhao

Social Network Analysis

Erica Seng-White

12/12/2020

SNA Research on Tweets Activities about Femeline Hygiene During Covid-19

Abstract

Feminine hygiene products can lead to many controversial problems such as high prices, lack of supply, and period poverty. Recently, more related news reports reviewing the related issues because of the pandemic amplified this social issue. For example, female doctors and nurses claimed unfair treatment during a pandemic because they received no supplies for feminine hygiene products. We want to find out if society has noticed those problems by analyzing Tweets. This essay will look into the change in people's reactions to hygiene products during the pandemic period. We will focus on the interactions on Tweets of the related topics. After we pull the data, we will construct a social network and perform the analysis with it. We will start by studying its five-number summary to identify the data set. Then we will use the Bootstrap method to resample the data and compare them with the T-test. Finally, we will compare the experiment with the hypothesis, yield the result, and give out suggestions for future studies.

Introduction

Recent news shows that many Chinese low-income families could not afford feminine hygiene products, especially during the pandemic period. There was a considerable shortage of tampons in early 2020 because they are not labeled as aids supplies that are female's necessities. All those observations existed for a long time but kept unnoticed. During the pandemic outbreaks, when social stability got questioned and chaos happened, social problems appeared more rapidly and more seriously than before. Furthermore, that is the current situation of feminine hygiene problems. However, if the problems showed on the news but received no attention, that means people have ignored the problem and chose not to solve it. By the end of the pandemic, society becomes stable as the problem, but the problem never gets solved and will eventually appear again. Thus, it is essential to determine if the public has listed this social phenonium as a problem. We will test it by studying Tweets.

We will use deductive research that we start with the hypothesis. We expected that because feminine hygiene problems become more severe than before, people on Twitter interact with this topic more often in quoting, replying, favoriting, and retweeting. Moreover, oppositely, if there are no reactions on related topics, that shows the public has limited awareness of the social problem. Thus, as the leading influencers, governments, mass media, and feminine hygiene product companies are responsible for guiding society to understand the problem. We will measure the reaction level by analyzing interactions in Tweets. If there were more interaction density on Tweets during the pandemic than before, we would say there were more general reactions. After performing the experiment and giving out the results, we will make

conclusions and suggestions for solving the problem based on how much attention people giving to the problem. We will also review the data collection and analysis process to make suggestions for future works of related research topics.

Backgrounds & Assumptions

The price and distribution problem of feminine hygiene products become more severe during the pandemic. This is also an essential premise for the theory because if there is no change in the problem during the pandemic, public reactions will not mean anything.

"You either have sanitary pads or a loaf of bread," stated the heading of BBC news. Because of the pandemic, there was a significant increase in the unemployment rate globally. The increasing number of people dropped below the poverty line and could not afford the primary lifesaving demand, such as food and sanitary pads. Moreover, in recent months, a post of "Loose-packed sanitary pad" has been on the top hot trend on Weibo for over weeks. The post shows a picture of a hundred unbranded sanitary pads sold only for four US dollars. Instead of questioning its quality, most replies complain about poverty and high prices for major sanitary pads products in a supermarket. Feminine hygiene products should be the necessary supply for everyday life, but most people cannot get or purchase them during the pandemic. News reported on this, and the public should be aware of this problem.

Methodology

Data Collection

The research will pull two sets of data with Rtweets. The first set will be the Tweets posted before the pandemic, and the second set will be the Tweets posted during the pandemic. By comparing the difference between the two sets of data, we can see which period has more interaction on Tweets. The population of tweets related to the topic would be massive, and there will be a limit for the maximum number pulls for Retweet. Thus, to ensure the limited sample pulled can maximumly represent the population, we need to plan the data collection process carefully. After comparing it with another method, we will use the ego with the alter method for the collection method. Because the total pulls are limited, we cannot perform a censure experiment.

Furthermore, because we need to randomly select the samples and set a boundary to the experiment, we will use the ego with the alter method. On the Twitter database, Retweet will randomly select several Tweets and pull every adjacent Tweet that has replied, favorited, quoted, and retweeted the selected Tweets. Rtweet will repeat this procedure until reaching the total pulls.

There will be two boundaries for network building, Tweets topic control, and posting time control. We need to extract only the related Tweets from the population. People post millions of Tweets every day about different topics. When collecting data, we will use the hashtag as the indicator to separate the population's data. Users often include the keyword as

hashtags into their Twitter. According to Twitter's website, the primary purpose of the hashtag is to categorize Tweets. A hashtag can be placed anywhere on a Tweet and typically include less than two keywords after the symbol #. Twitter can push the related topic for users based on their interest and calculate and rank the hot trend topics (Twitter Help Center).

On the other hand, users can quickly search for Tweets based on hashtags. Moreover, by adding the hashtags, User's tweets have a greater chance of showing on others' main screen, which is often used by influencers. Social issues can be determined by the hot trend keywords and hashtags of Tweets. If the number of tweets shared the same hashtags exceed a certain threshold, Twitter would list the topic as a hot trended topic, such as "#CaliforniaFires" or "#Halloween."

In this experiment, we will pull Tweets that contain the hashtag "Feminine hygiene." The purpose of choosing this hashtag is to prevent false negative measuring error. If we use other hashtags like "Sanitary pads" and "Tampons," we may exclude some products' hashtags, so we exclude the nodes and edges that should be included. This false-negative measure will damage the collection between nodes. For example, if the research does not contain the hashtag "menstrual cup," collections of nodes will be missing and underestimate the network's density.

We will collect two sets of data containing hashtags of "feminine hygiene" and within two different time boundaries. The first set of tweets covers April of 2019, and the second set of tweets covers April of 2020. We want to control dependent variables by pulling Tweets for the same month and different years. For example, we can prevent other factors from influencing

these hashtags' popularity, such as seasonal differences. The outbreak of pandemic happened in March in the United States. However, we choose to study data from April because, after one month, there would be a shortage of medical and market supply of feminine hygiene products. If that situation happened, there would be more news related to the shortage problem, and people would mention more on social media, including Twitter. Thus, we can test the hypothesis that there were more Twitter reactions, including posting, replying, quoting, and favoriting Tweets in April 2020, than in 2019's.

Network Structure

Nodes, attributes, and edges are the fundamental characters for social networks. In this research, we study the connections between Tweets to use individual Tweets as nodes. The type of nodes will be unipartite, meaning there is one kind of node for the networks. The interaction between nodes will be the edges. We want to include all kinds of interactions, so there will be multiple types of edges, including favoriting, replying, quoting, and retweeting. We believe that all four activities mean people are interacting with others using Twitter. For example, if a user agrees on a specific idea, the User is more likely to favorite or retweet that Tweet. Furthermore, if a user disagrees or wants to add comments, that User is more likely to reply to the Tweet. Therefore, we will combine all four activities when connecting the nodes. On the R script, when converting the data list to the graph, we will use e=" all" to include all activities.

Similarly, when measuring the power of a single node, we will measure its degree of centrality. The more connections that node has, the more activity the Tweets owns, which means that Tweets is the center of the discussion. For instance, Tweets posted by the influencers are

likely to be the hub on this network because there will be more retweets and favorites. This power will also be the main attribute when plotting the network. If a node has more connection, it has more power, and the size will be larger than other nodes.

Resampling & Comparing Methods

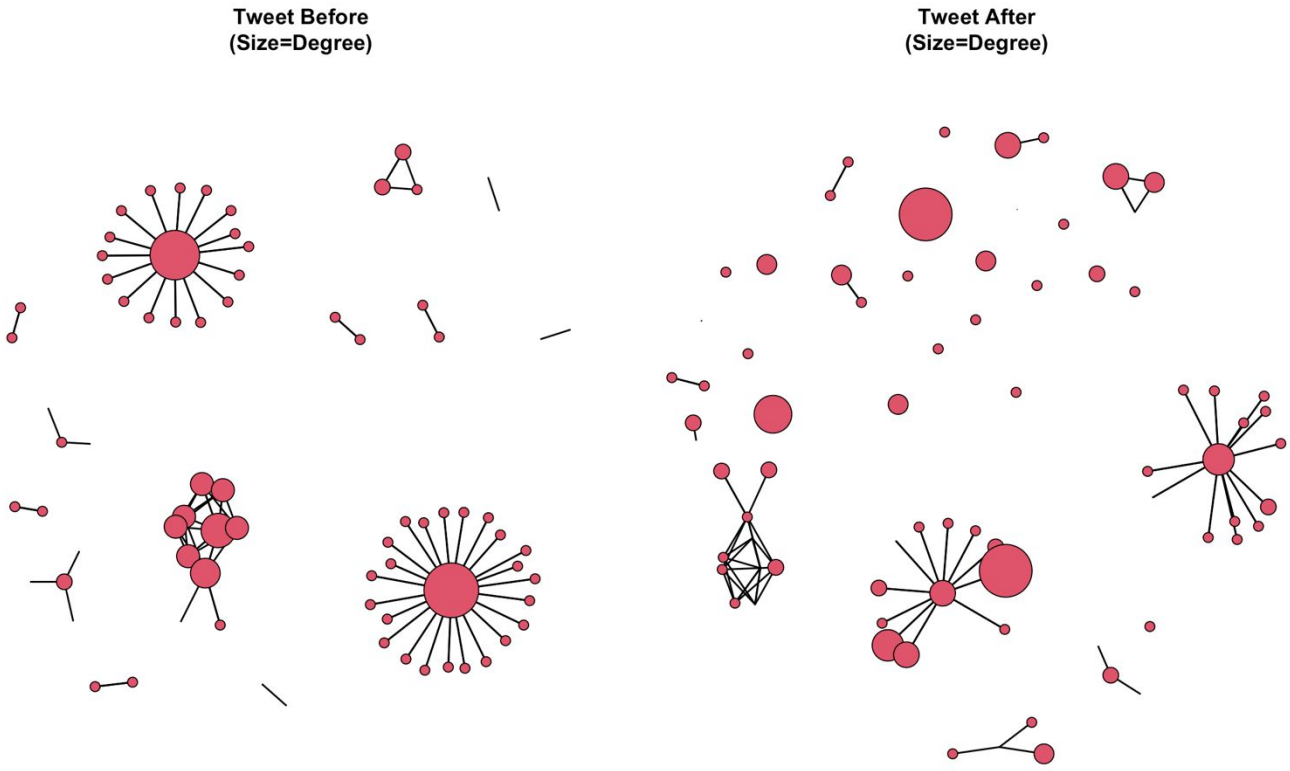
we will use the Bootstrap method to reassemble the data and get the most accurate standard error with our data's limited sample size. Then we use the T-test to compare two sets of data to find out if the result is statistically significant to reject the null hypothesis. The Bootstrap is a resampling method that allows us to treat a particular portion of the data as a network population. By using Bootstrap multiple times, we create a null distribution for both networks. In this research, we repeat the Bootstrap process 1000 times, and for each Bootstrap test, we measure its density and merge them to the density lists for each two data sets. With the density lists, we can calculate the standard error between the sample and the whole population. For this research, Tweet pulled the sample, and the population is all Tweets with the hashtag "feminine hygiene." Standard errors can show the differences between the calculated density and the mean of the density. We expect to see a significant difference between the density from the two data sets (0.05 cut point in P-value and 1.65 cut point in T-value.) Therefore, the null hypothesis becomes that the reaction in 2020 is less or equal to the response in 2019, and there is no significant difference between the two sets.

These two data sets are longitudinal because they measure the activities for Tweets in different periods. However, the two data sets' standard errors are different because the research population in 2019 is different from the population in 2020. For example, users who tweet about

feminine hygiene products in April 2019 would not tweet again in April 2020. Thus, there will be various standard errors for the two data sets, and we use an independent T-test to compare those two networks.

Network Plots

For quantitative network research, position and size are two of the best attributes when plotting the network. As a result, We use the Force Directed Algorithm, also known as distance scaling, to determine nodes' position. This algorithm pulls nodes with more ties closer together and nodes with fewer ties further apart. For the Tweets networks, this algorithm will help distinguish groups and individual Tweets. Moreover, the activity attributes the difference in one node will show in the size difference, so the more extensive the size means more activities on those Tweets.



Data Analysis

Five Number Summary

This part of the report summarizes two networks: size, density, components, diameter, and clustering coefficient. For the size characteristic, we set the data sample size for both sets to be 100. We collected a full 100 Tweets for the April 2019 (Before) data set, and we were able to collect 80 Tweets for April 2020 (After) data set. This research population is every Tweet posted during April of 2019 and 2020 with the hashtag "feminine hygiene." Moreover, the samples are the Tweets that were randomly pulled by Rtweet. We use the ego with alters method for pulling Tweets, and with the limited quantity pulls, there are advantages and biases for the research that will show on the final suggestions.

Density is another primary characteristic for a social network, and it represents the proportion of actual edges to the maximum possible advantages. In this research, edges are a combination of mentions, quotes, replies, and retweets. There are 77 mentions, seven replies, and 50 retweets on the before data set. Furthermore, there are 54 mentions, four quotes, five replies, and 35 retweets on the after data set. We use the combination to calculate the density because they represent the frequency of interactions on Tweets. If a user mentions, reply, quote, or RetweetRetweet a Tweet, it shows that the User has an agreement with the original Tweet and shows the importance of that Tweet. The density of the before data set is 0.013, and the density of the after data set is 0.015.there is a slight difference of density between the two data sets and support the hypothesis. However, it is only a parameter and cannot provide strong evidence to test the hypothesis. We need a T-test for a network with a large amount of data and an unknown standard deviation.

Quotes, retweets, replies, and favorites will all form components, and they suggest the interactions between Tweets. There are 37 components for before the data set and 40 for the after data set. Most of those components have less than three nodes, but three major components have multiple nodes for both data sets. Thus, two networks have only slightly different component numbers. Diameters of two networks are similar as well, 3 for the before and 2 for the after. Diameters illustrate the largest number of the shortest path between nodes. In this research, we expect small diameters because the relationship between Tweets is often binary. Finally, both networks' transitivity is one because it is rare that all three Tweets retweet or quote each other at the same time.

Resampling & T-Test

We first convert both Starnet data sets into an adjacency matrix for Bootstrap. Then we conduct two resampling processes 1000 times and get 1000 simulated matrices. We calculate the density for those data, convert the matrix back to start, and get edge density for every simulated data set. With the function of standard error, we get the standard error for before and after data set: 0.0059 and 0.0075. With the calculated density and standard error for both data sets, we can perform the T-test for them.

The equation of the independent T-test is $(D-D)/\sqrt{(SE^2)+(SE^2)}$, where D represents the density of two data sets and SE represents the standard error of two data sets. The value of the T-test result is -0.205. The smaller the T, the smaller the evidence against the null hypothesis and with less significant differences. In this case, the T-test result's absolute value is 0.205,

which is less than 1.65 showing there is no significant difference, and we cannot reject the null hypothesis. The result is against the theory, and there is less difference in reactions on Twitter over the pandemic time.

Conclusion & Future works

There are two primary findings from the result of the T-test. Because the result is a negative number, it shows the activities on Twitter after the pandemic is more than the activities before. However, the difference cannot reach a significant difference. Therefore, the result shows that the public has equal awareness of feminine hygiene issues during the pandemic. Without addressing the problem, we can never solve the problem. As a result, governments, mass media, and other influencers must guide society to realize this problem.

We also find out the possible bias during the research that could cause or Type II error. During data collection, there is a possible false-negative error. We set the boundary for the population as Tweets with the hashtag of "feminine hygiene." However, many other hashtags are excluded but related to the subject, such as "tampons" or "sanitation cup." The sample size is insufficient. With a maximum of 100 data on each set, we can only observe approximately three large components with activities, so more sample size must reflect the population. Plus, when we pull data from Twitter, the program uses the ego without an alter method. With the limited sample sizes, we cannot perform the full-scale network measurement like distance and centrality. In this research, information for the outer layer of the component will be missing and decrease the network's density. For example, Tweets from the outer layers have their mentions and replies, but the data pulls end at that point, so their mentions and replies are not included in the research.

There are some alternative research methods, as well. For this research, we focus only on the density of all four activities. Future research can use ERGM to study possible adjacency matrices on a specific attribute. For example, we can test the network's change if we only study the replies and retweets.

Reference

BBC news. (2020, December 4). Covid poverty: 'You either have sanitary pads or a loaf of bread', Retrieved from https://www.bbc.com/news/uk-england-leeds-55150671

Bharathi, SP. (2020, May 3). Unable to afford sanitary pads, more women depend on volunteers for supply, Retrieved from https://www.thenewsminute.com/article/unable-afford-sanitary-pads-more-women-depend-volunteers-supply-123856

Feminine Hygiene and Sanitary Protection Products: Incl Impact of COVID-19 - US - September 2020. (2020). Mintel Group Ltd.

Twitter. Help Center. Retrieved from: https://help.twitter.com/en/using-twitter/how-to-use-hashtags