Analyzing influencers and geographical sentiment distribution for tweets on the Zika Virus in the United States

Gade Venkata Sai Akshay (Team leader)
G01025094
Data Analytics Engineering
Vgade2@gmu.edu

Saurabh Rao Donthineni
G01025113
Data Analytics Engineering
sdonthin@gmu.edu

Kaushik Kandlakunta
G01026145
Data Analytics Engineering
kkandlak@gmu.edu

Abstract- The zika virus has been a major source of concern in the United States. People are using tools like Twitter to gather opinions on the virus, help them make informed decisions and express their opinions on topics related to the issue. Through this project, we aim to understand the major keywords that are used by users tweeting about zika, then classify those tweets by sentiment and visualize the sentiment geographically over time. We also aim to understand the major influencers and news makers who determine the direction in which the conversation about the zika virus is heading.

Index Terms – Zika, Social media, Twitter,United States of America

I . Introduction

Social media data is a very good indicator of how events are transpiring in the world. It is possible to get a very real sense of how information is disseminating across the globe by analyzing this data. Not surprisingly, the trend of using social media data for information access has spread to the field of healthcare as well [1]. These trends are collectively referred to as the "medicine 2.0" trends and broadly encompass the usage of web tools including blogs, geotagging, podcasts, wiki entries etc. These are used by all major stakeholders in the healthcare industry, including doctors, patients, administrators, to open source collaboration with the intent of personalizing healthcare services and promoting healthcare education [2]. The study is to help stakeholders understand the kind of information that is being shared with regard to the Zika virus, and who are the major influencers that are driving the conversation. This would help government and private entities with the assessment of Twitter messages by killing the unsustainable procedure of physically checking a huge number of tweets after a Zika outbreak , to identify potential users who can help with the dissemination of accurate information online. The point of the study is along these lines not to concentrate on sifting through superfluous online networking redesigns, but rather the examination plans to build up a strategy that imitates a manual human assessment process utilizing an arrangement of robotized strategies that lessen the unmanageable number of tweets to a sufficiently little specimen that can be promptly surveyed by the emergency services for basic noteworthy data. We also aim to understand the geographical distribution of the tweets to see the amount of content that is being generated in each region.

II Related work

While there is a lot of user generated data on twitter where people express their opinions, there have not been a lot of efforts to fully understand their opinions and harness it to power crucial business decisions. Our work has been motivated by several investigations that have occurred in this space. Scanfeld, Daniel, Vanessa Scanfeld, and Elaine L. Larson [4] have investigated the dissemination of health information through social media networks, specifically, twitter. They did confine their investigation to data that related to the term "antibiotics". Nastasi, A., Bryant, T., Canner, J.K., et al. (2017) [5] focused on how hashtags generated by users can be used by healthcare providers to understand their potential market. We wanted to work on a unified solution that would help both government and private businesses understand the geographical dissemination of healthcare information to ensure that they can better direct their resources to focus on areas that need more attention. Salem, J., Borgmann, H., Bultitude, M., Fritsche, H. M., Haferkamp, A., Heidenreich, A., … & Tsaur, I. (2016) [6] did make a compelling case for using geolocation data to drive the conversation regarding the spread of information geographically through social media. Additionally, there has also been some research done to track the spread of ailments over time, which is especially important when the disease being considered can spread over larger regions over time (for instance, swine flu at the macro level and common cold at the micro level). Paul, Michael J., and Mark Dredze [7] focus on approaching this problem by using syndromic surveillance to track the spread of disease over time. Additionally, the paper also focuses on localizing illnesses by geographic region.

III Key questions

This team has identified a set of key questions that can be answered. Rather than simply focusing on the

technical aspects of the project, we believe that answering this question lends the project some credence regarding its applicability in real life. Given the very broad nature of microblogging users, it might not be possible to answer these questions.

1 Who are the major influencers who are driving the conversation about Zika in the United States ? When it comes to tweeting about issues of health, corporate entities , individuals and disease specific accounts have a large share of the tweet population. We believe that individuals are the ones driving the conversation about Zika, and we aim to find the influencers, if any. Corporate entities and disease specific accounts might also account for a large percentage of the tweets, which is what we aim to understand. We also aim to understand the relationship between the accounts that are tweeting about the Zika virus, and understand how the communities are connected.

2 What is the geographical distribution for the tweets like ?
Are people on the East coast or west coast more concerned with the spread of the Zika virus. We expect people on the East coast to be more concerned about the Zika virus than those on the West coast, simply because the number of reported cases have been higher in the aforementioned region. For the study, we have considered the ten major regions in the United States which have the highest population. The major reason behind doing this was that these regions have a larger percentage of people who are likely to tweet about Zika. Additionally, since we would be considering almost 19% of the population of the United States of America, we believe we will have a representative sample of how the conversation about Zika is structured. We expect cities with major airports that have a higher chance of receiving people that have been infected with the zika virus to be more concerned about the virus as such, and expect a larger percentage of tweets from them.

3 What is the overall sentiment of the tweets

Zika happens to be a disease that has not seen a lot of cases in the United States of America. The number of reported cases has been small, and has oftentimes been contained. This is the major reason why we expect the tweets to be informational in nature, and to therefore have a neutral sentiment. We do not expect tweets to focus on emotions , as this is not an issue that is currently being faced by people.

IV Proposed Approaches

Our early approach focused on :

1. Understanding the major keywords used to talk about zika data, and the distribution of said keywords over time to understand the growth and retardation of topical issues over time. The major purpose of this is to understand what issues are being talked about in general and to understand the most common topics regarding zika that people are talking about. During this process, we also aim to understand the network of influencers who do drive the conversation and influence opinions about healthcare on social media data. The idea is also to understand a network map of the influencers so this can be visualized easily.

2. Geographical distribution – to geographically distribute the tweets which have been posted to understand which regions have a higher focus on zika. This would help government and private entities decide how and when to allocate their resources to ensure the maximum chances of survival for the affected patients. Additionally, we also want to understand what the overall sentiment is over a fixed time delta.

1 Data Source : For this exploratory research, we used data from Twitter. Through the use of hashtags, tweeting, and re-tweeting, it allows users to voice their concerns, connect with others, and show solidarity, while also demonstrating their person and/or organizational identities. We gathered tweets that had the term “Zika” in them for a seven day period, using the twitter REST API. To ensure that we were gathering data from each region, we also specified the latitude and longitude value for each region that we felt was an accurate representation of the population densities they harbor. We chose to find all tweets within a 100 mile radius from each latitude and longitude value.

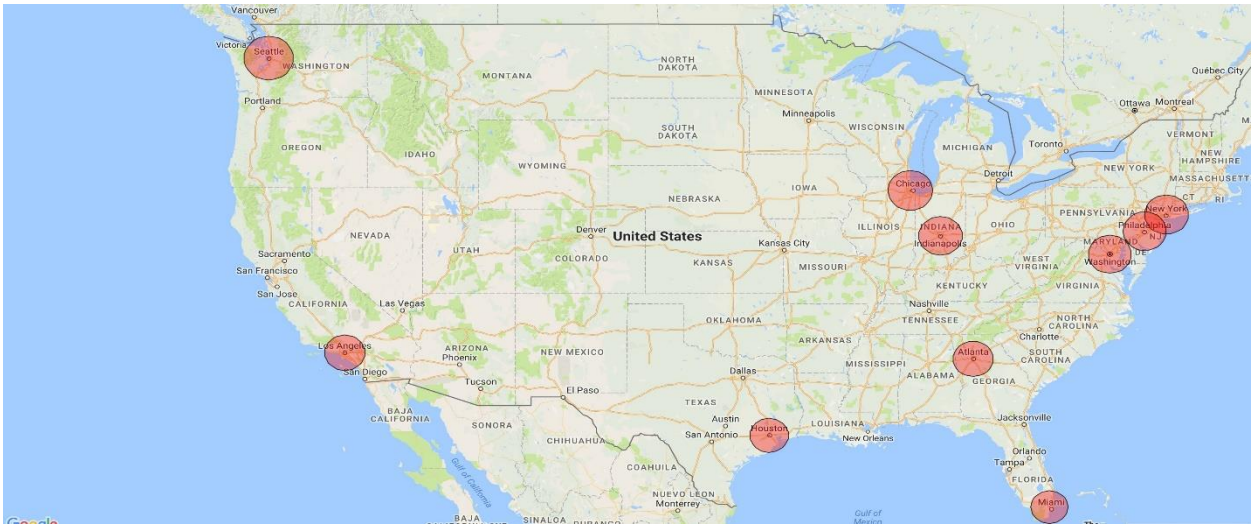
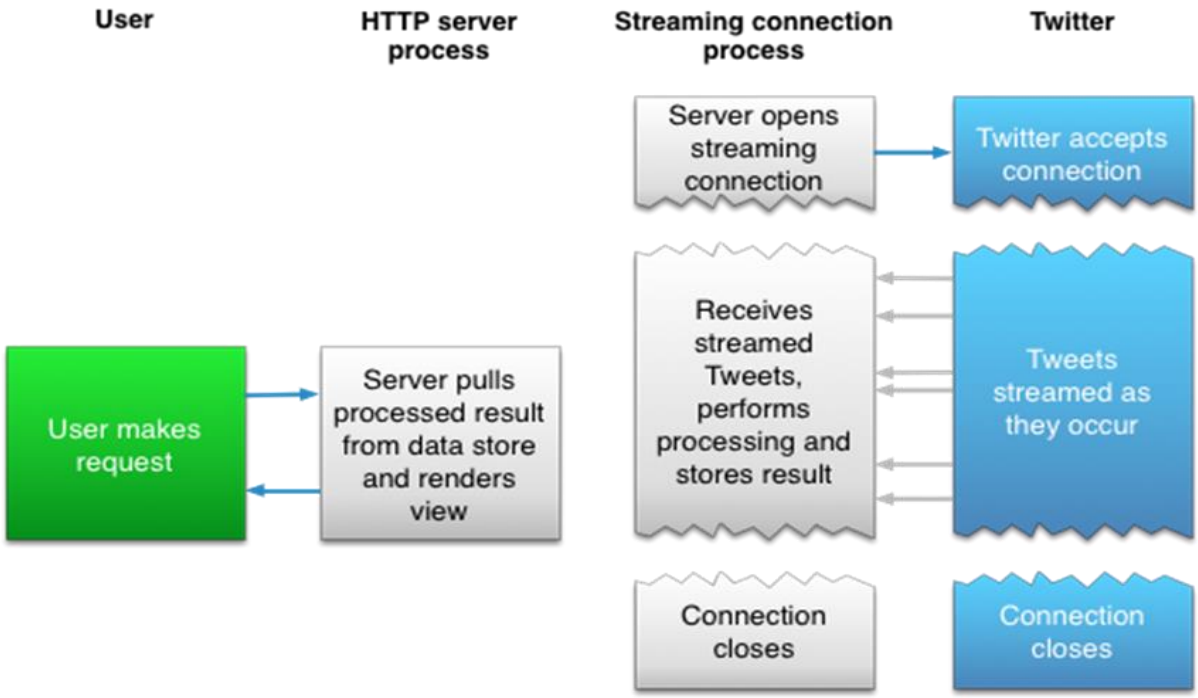


Fig: Geographical distribution of the cities that have been considered for the analysis

V System Design

A) ARCHITECTURE The tweets are mined from twitter using their Streaming API. The data gathering is done using twitteR package in R. We used the REST API that Twitter offers to gather tweets. Twitter allows gathering pre-indexed tweets using the REST API. Developers are also allowed to specify different parameters in their search query , including parameters like start date , end date , type of language of tweets , location information etc. The process involves the user making a request , after which the server pulls the results based on the search parameters and renders the view , which can then be locally saved o the disk after receiving it as a json stream.



B)

Fig : A graphical explanation of the REST and Streaming APIs offered by Twitter.

1 Data description

We chose to consider 16 variables for each tweet, from ~45 to 50 variables that a tweet has. Since the major purpose of this happens to be to understand how to model the network graph, we focused on gathering close to 16 variables, which we used to gather the source and target nodes for the network graph and also to assign the weights for the edge graph. The variables that were considered and their data type were :

1. Created – time and date variable
2. favoriteCount – integer
3. Favorited id – long integer
4. isRetweet – Boolean (T/F)
5. Latitude -long integer
6. Longitude – long integer
7. replyToSID – string
8. replyToSN – string
9. replyToUID – string
10. retweetCount – integer
11. Retweeted – Boolean (T/F)
12. screenname – string
13. statusSource – string
14. Text – string
15. truncated - Boolean(T/F)

2 Data cleaning

After we obtain the tweets in their raw form , we save them to disk in the form of an comma separated value file. We load this file into R and further process it to clean the data. We remove the stopwords and change the character encoding to UTF-8 from ASCII, which ensures that there is a machine readable format for the sentiment analysis. The major reason that the character encoding conversion is being done is because Twitter allows users to post tweets that have emoticons and characters that are not a part of the UTF-8 encoding scheme. After this is done, we remove the stopwords and convert each tweet to lowercase. We do this for better processing during the sentiment analysis stage of the project.

3 Databases

A database management system is vital in light of the fact that it oversees information proficiently and enables clients to perform various assignments effortlessly; this is an extremely normal accessible meaning of what a database is. What's more, here, we plan to use the database for its definition and enable our extend with a full time devoted database to help us utilize cross stage method to accomplish our objective. For this project, we have saved all of the data we need to a comma separated value file stored on the local disk.

4 Exploratory data analysis

For the exploratory data analysis , we wanted to understand how the tweets are distributed geographically plotted against their populations. For the regions we considered, the total population was 61,125,196 , which is 18.92% of total US population. The total tweets we have are 3095. We divided the tweets from each region and the population from each region by the total number of tweets and the total population to get a fraction of the users from each region.

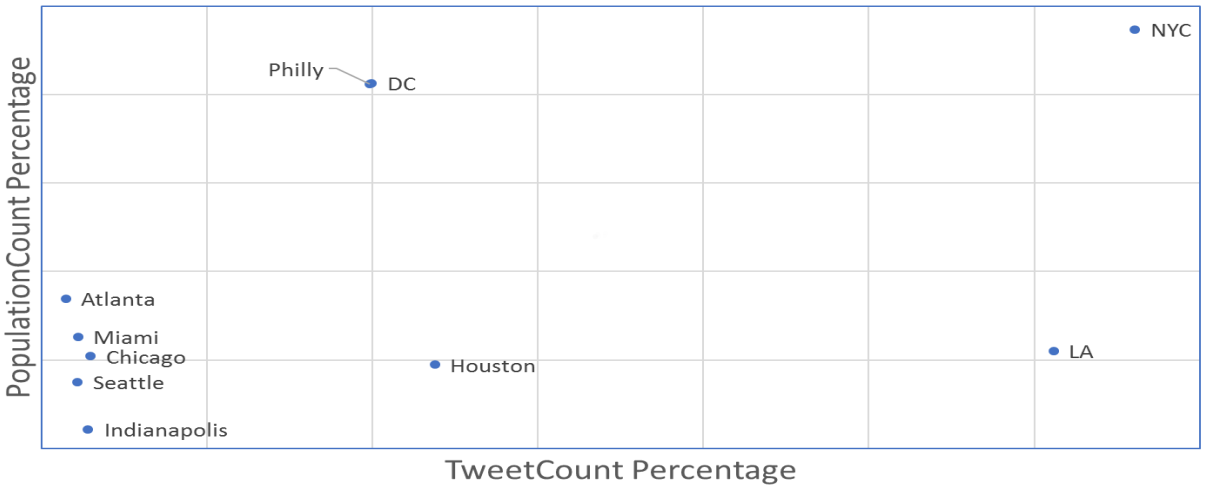


Fig: Tweetcount percentage Vs PopulationCount percentage

There are some very interesting observations here. New York city has the highest population , and the number of tweets that it has is also proportional to the population it has. Los Angeles on the other hand has a very large population , but the number of tweets that is has is not proportional to the population. Philidelphia and Washington DC have almost the same kind of plot on the x-y axis, which can also be explained by the fact that the regions are very close, so the kind of users that will exist in both these regions will be same. We also plotted a basic wordcloud to understand the major keywords that are being used to talk about the Zika Virus.

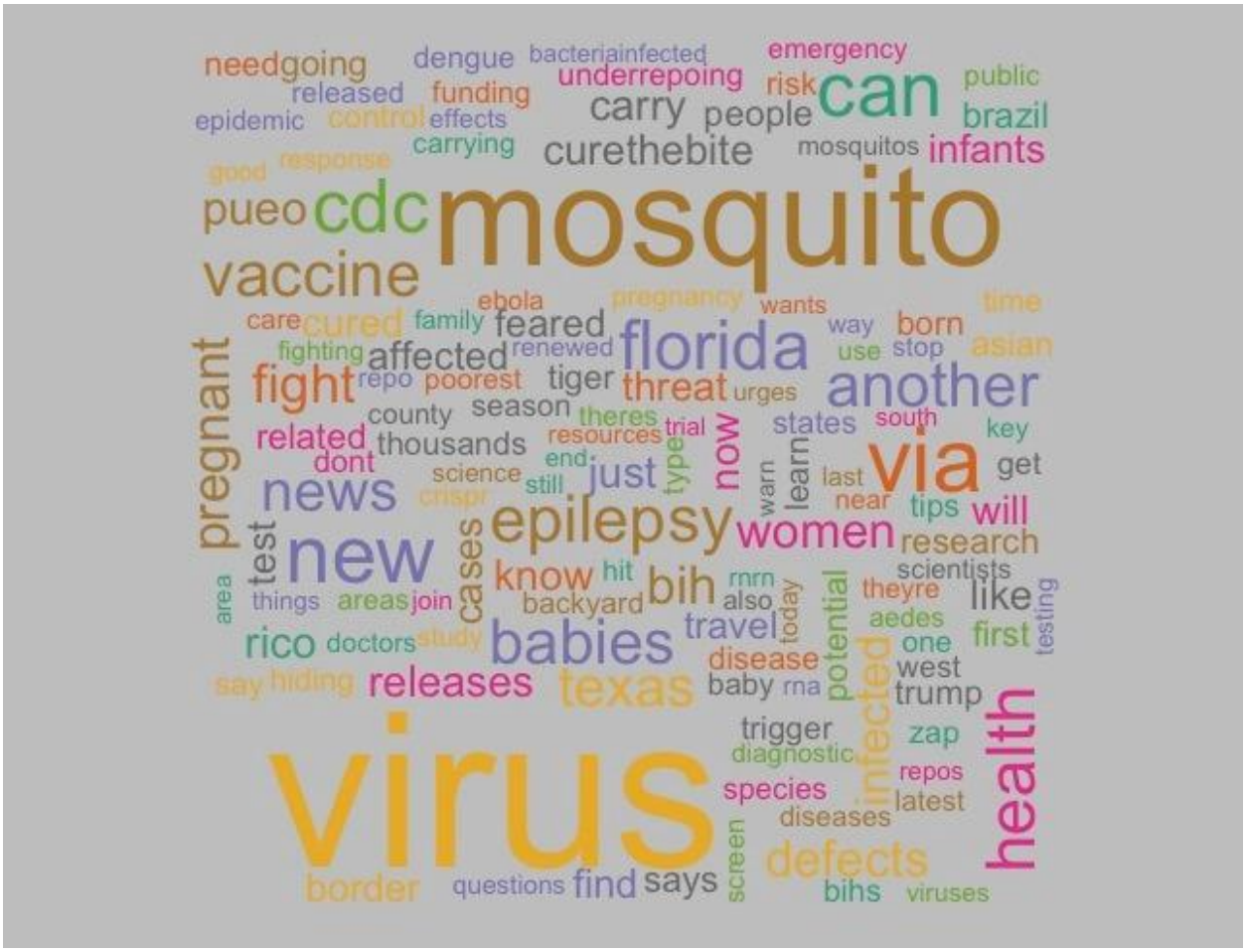


Figure : Wordcloud

5 Network Modelling

We wanted to understand the relationship between different users and also who is more influential. The characteristics of social networks allowed them to be modeled to understand the propagation of information through them. The network would primarily focus on modelling nodes and edges for the graph , and the weight of the edge being decided by a combination of the tweet's retweet count , favourite count and whether the tweet is a reply or not.

6 Visualization

For modelling the visualizations , we used Gephi and Javascript. We used Gephi to first model the initial node and edge list. After we understood how the graph looked like , we also made worked on an interactive visualization using Javascript, which would import all the functionality that Gephi has to offer with regard to node edge values and plotted them in the interactive visualization. We used Javascript to make the graph clickable.

B) LANGUAGES BEING USED

R – to mine tweets from twitter using Twitter's REST API.

Python – for data pre-processing

C) SOFTWARES TOOLS

Import.io

Microsoft Excel

Tableau

D) DATA SETS

The data will be gathered from Twitter using their REST API.

VI Experimental results

1 Sentiment analysis

For the sentiment analysis , we consider a corpus consisting of approximately 7000 positive words and 7000 negative words. For each tweet , we calculate the total sentiment score by first calculating the positive sentiment score and then calculating the total negative sentiment score. The positive sentiment score is a

positive value and the negative sentiment score is a negative value. We then calculate the total score by adding the positive and the negative score. After this is done , we assign a label of positive , negative or neutral based on the overall score for each tweet.

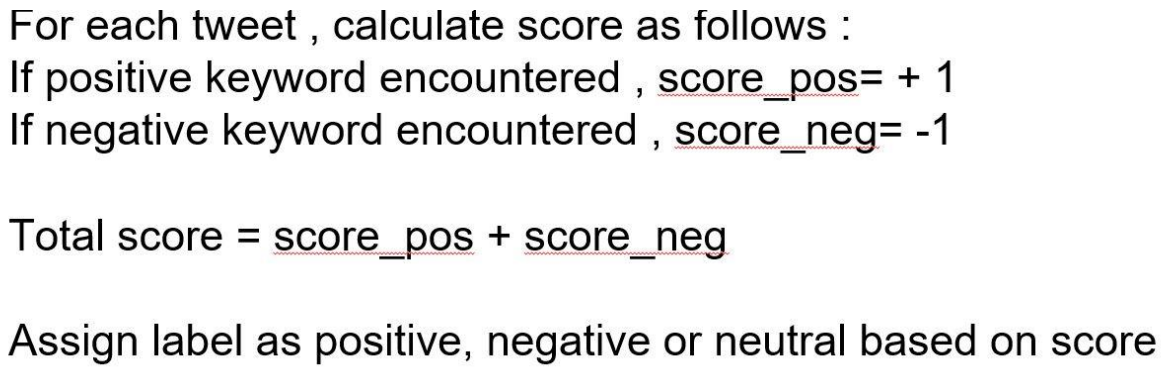


Fig : Sentiment score calculation process

We repeat this process for each geographical region and calculate the sentiment distribution over time.

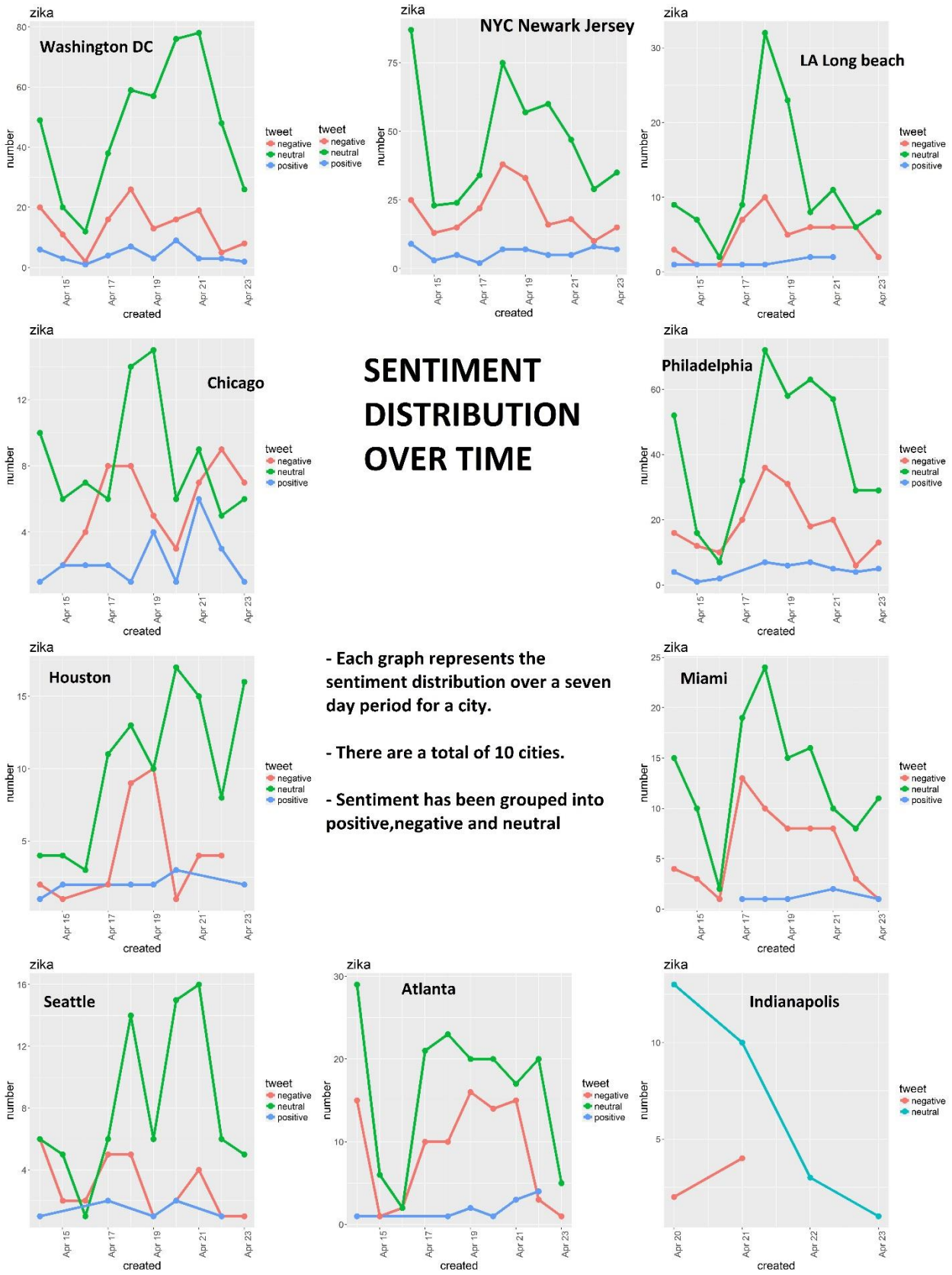


Fig : Sentiment distribution for each region that has been considered.

2 Graph analysis

For the graph analysis , we have to first calculate the weights for each edge node pair. For this , we have considered normalized reteweet count , normalized favorite count , reply or not , where the first two variables are between 0 and 1 , and the last one is 1 or 0 .

First , we normalize the retweet and favorite count by dividing by the maximum value for them. Then , we assign a value of 1 or 0 to a tweet based on whether it is a reply or not. Then , we calculate the weight by using the formula weight= X1*(Normalized_rt_count) + X2*(Normalized_fav_count) + X3*(Reply_or_not) ∋ for tweet_count=(1:n) .

The values X1 …X3 are constants and are heuristically determined. This is done by calculating the mean , median and mode for the weight values and ensuring that mean=median=mode for ensuring the best possible normal distribution.

Considered retweet count , favorite count , Reply or not for weight.

Normalize retweet count, favorite count by dividing every entry by maximum value in each column (ie, maximum possible value is 1).

Reply or not is a Boolean value , convert to 1 for TRUE , 0 for FALSE.

Assign weights by using the formula weight= X1*(Normalized_rt_count) + X2*(Normalized_fav_count) + X3*(Reply_or_not) ∋ for tweet_count=(1:n)

{ Σ(Normalized_rt_count[tweet_count] + Normalized_fav_count[tweet_count] + Reply_or_not[tweet_count]) = Σ weight }

Where X1,X2,X3 are constant values ∋ 0 ≤ X1,X2,X3 ≤ 1 . X1,X2,X3 are heuristically tweaked to ensure the optimal degree of normality. Condition for normality : mean(Σweight) ≈ median(Σweight) ≈ mode.mult(Σweight)

Figure : Process of calculating weight

We get three graphs , one with only retweets , one with only replies and one with both of them . The graph metrics for each of them have been mentioned :

Variable	Retweets	replies	retweet+replies
connected components	1.311	1.229	1.332
avgweighteeddeg	0.099	1.269	0.274
network diameter	3	1	4
avg path length	1.34146341	1	1.4297082
density	0.001	0.004	0.001
modularity	0.629	0.821	0.871
modularity with resolution	0.629	0.821	0.871
num of communities	375	76	427
pagerank			
epsilon	0.001	0.001	0.001
probability	0.85	0.85	0.85
connected components			
no of weakly connected components	359	71	397
no of strongly connected components	984	166	1118
average clusterting coefficient	0.003	0.004	0.004
eigenvector centrality			
no of iterations	100	100	100
sum change	0.0638098	0.01096035	0.084305431

Fig : Graph values for each type

Overall , we can say that :

- Most conversations about Zika in the United States are influenced by corporations and issue specific accounts rather than individuals.
- Regions close to the East coast are more concerned about the Zika virus than those in the Midwest , center or West coast.
- The overall sentiment is neutral, owing to the informational nature of the tweets.

3 Graph parameters

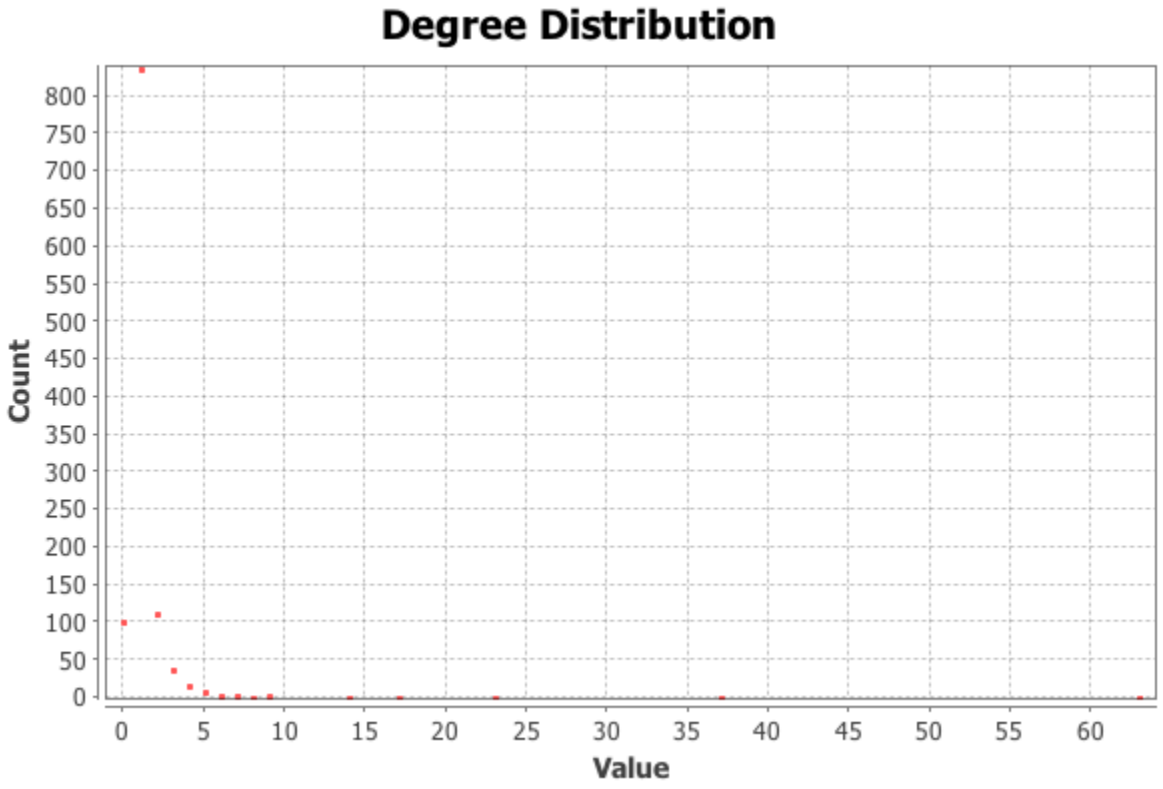


Fig: Average Degree Distribution

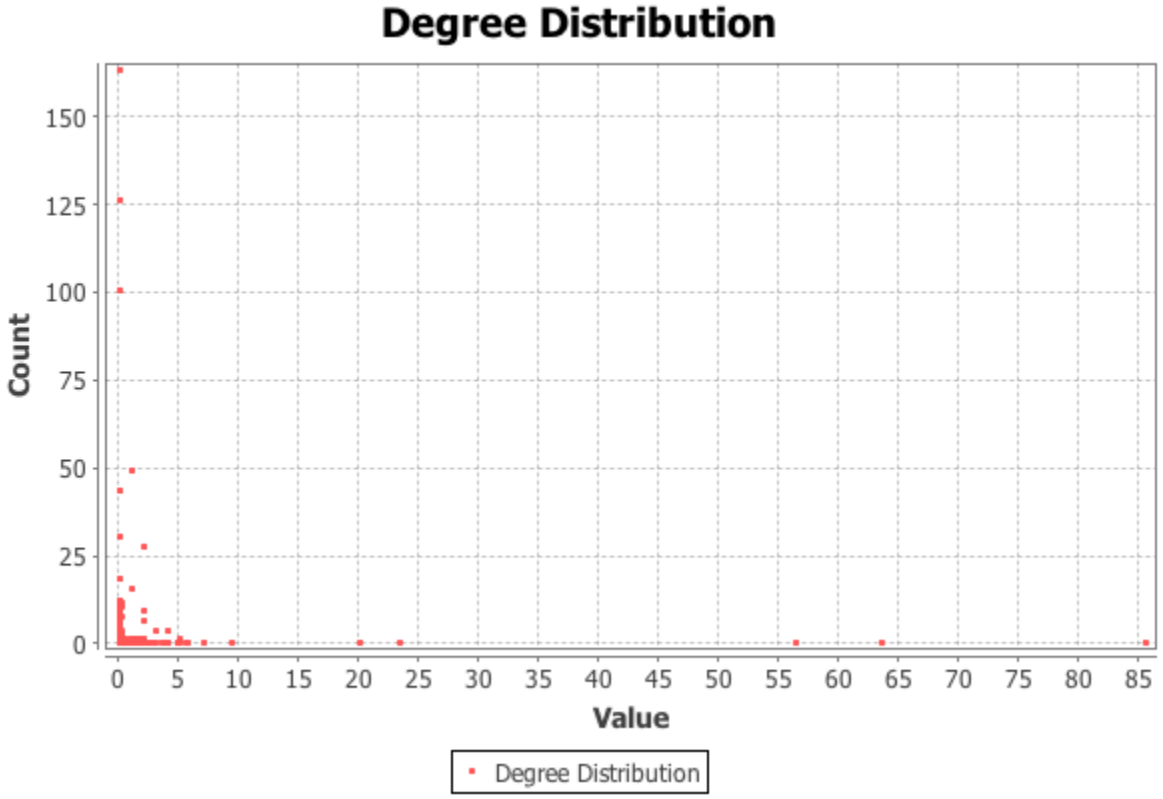


Fig: Average Weighted Degree

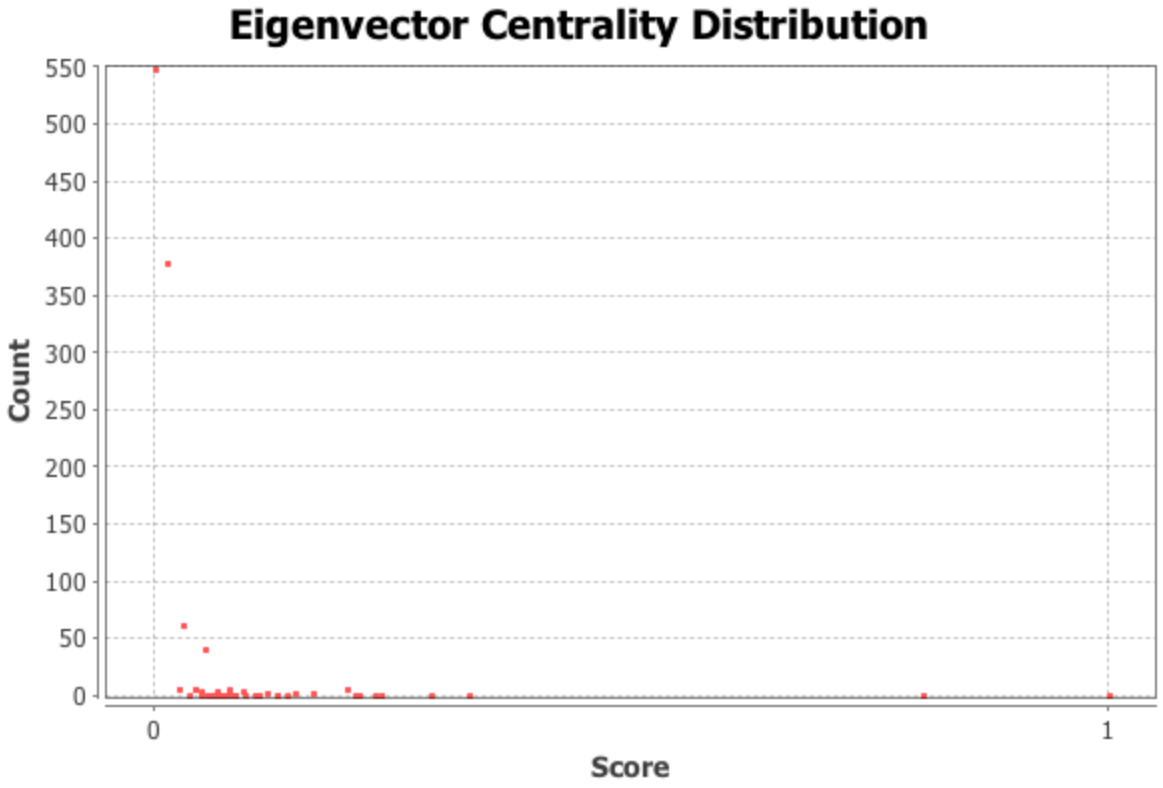


Fig: Eigenvector Centrality Distribution

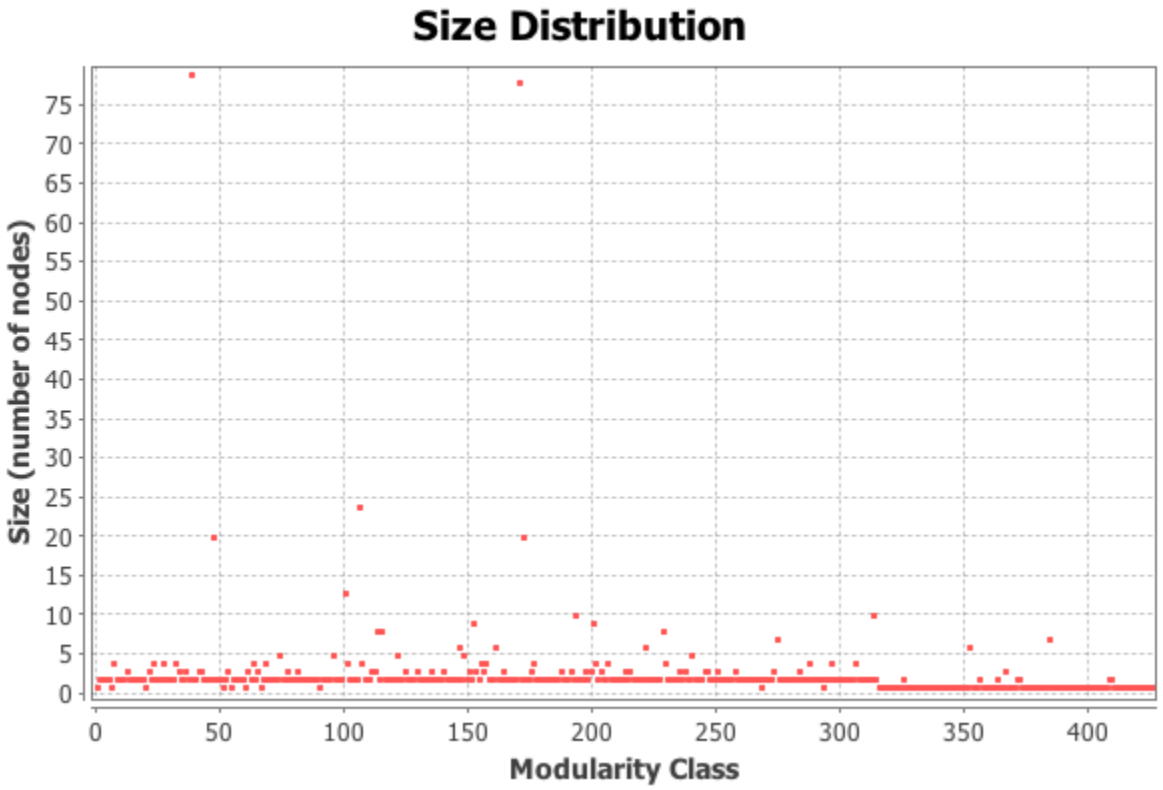


Fig: Modularity

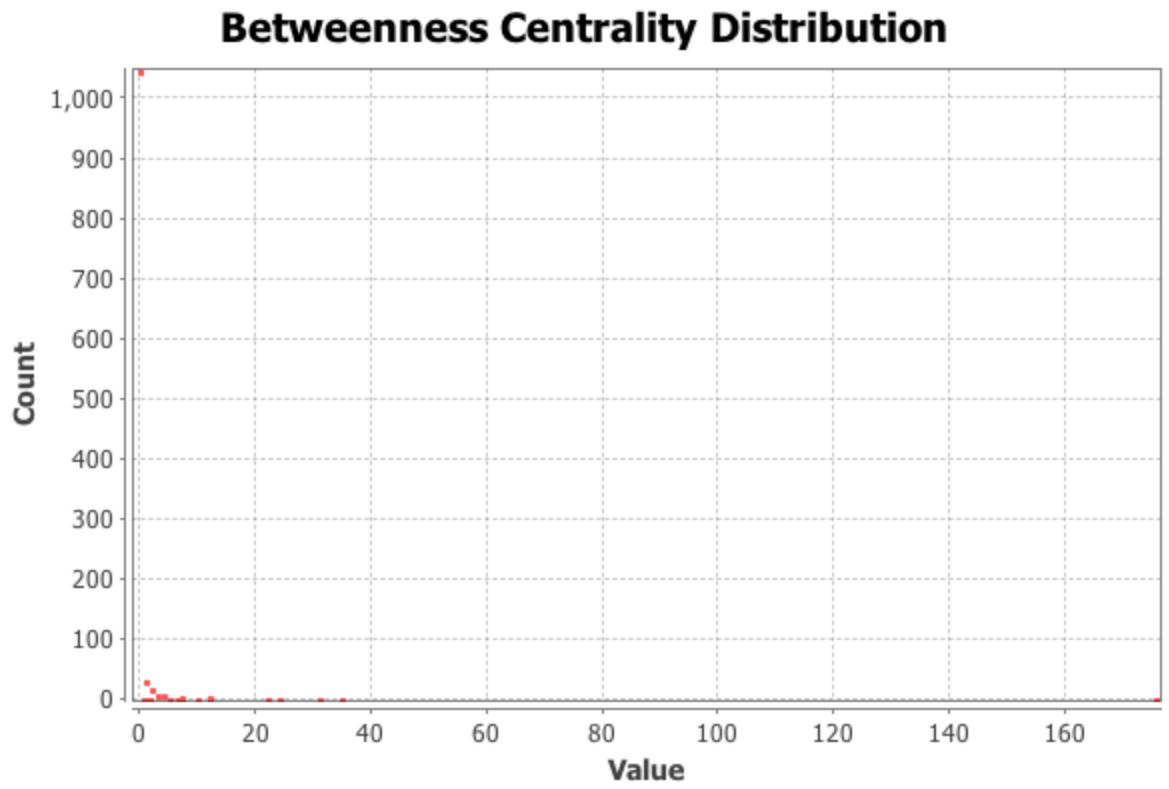


Fig: Betweenness Centrality Distribution

VII Limitations

The keyword we considered was a single one , and this might just mean that people are using this word with a higher frequency. In other words , there is no contextual understanding of the tweets, which can be improved.

VIII Future Work

Instead of using a REST API , we can use a Streaming API to understand how the networks are interacting in real time. We could also focus on tf-idf calculation for better understanding the keywords being used.

IX.APPENDIX

APPENDIX 1 – Databases and cleaning: Gade Venkata Sai Akshay EDA and mining: Saurabh Rao Donthineni Visualizations: Kaushik Kandlakunta

X References

X.REFERENCES 1. Trust evaluation in health information on the World Wide Web. Moturu ST, Liu H, Johnson WG Conf Proc IEEE Eng Med Biol Soc. 2008; 2008():1525-8.

2. Health 2.0 and Medicine 2.0: tensions and controversies in the field. Hughes B, Joshi I, Wareham J J Med Internet Res. 2008 Aug 6; 10(3): e23.

3. Fox S, Jones S. The Social Life of Health Information. Pew Internet & American Life Project. 2009 4. Scanfeld, Daniel, Vanessa Scanfeld, and Elaine L. Larson. "Dissemination of Health Information through Social Networks: Twitter and Antibiotics." American journal of infection control 38.3 (2010): 182–188. PMC. Web. 24 Feb. 2017.

5. Nastasi, A., Bryant, T., Canner, J.K., et al. (2017). Breast Cancer Screening and Social Media: a Content Analysis of Evidence Use and Guideline Opinions on Twitter. Journal of Cancer Education, pp 1–8.

6. Salem, J., Borgmann, H., Bultitude, M., Fritsche, H. M., Haferkamp, A., Heidenreich, A., … & Tsaur, I. (2016). Online Discussion on# KidneyStones: A Longitudinal Assessment of Activity, Users and Content. PLOS ONE, 11(8), e0160863.