Understanding the geographical distribution of healthcare focused tweets and measuring sentiment over time

Gade Venkata Sai Akshay (Team leader)
G02013456
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- **Healthcare data on social media has been proliferating with a large increase over the past few years. People are using social media tools like Twitter to gather opinions on healthcare providers, help them make informed decisions and express their opinions on issues in healthcare that concern them. Through this project, we aim to understand the major keywords that are used by users tweeting about healthcare related issues, 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 healthcare is heading.**

Index Terms **: healthcare, data, social media, twitter, geolocation, sentiment**

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, for the purpose of open source collaboration with the intent of personalizing healthcare services and promoting healthcare education [2]. The study is to help emergency services to distinguish healthcare data on online networking utilizing singular client tweets and upgrades. This will 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 disease outbreak or to track the spread of medical conditions in general, to rather center assessment endeavors on a modest bunch of focused messages with the most astounding level of pertinence. 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 aim to answer the following questions :
1. What are the major keywords that can be used to characterize the spread of a disease ?
2. How is a disease spreading geographically over time ?
3. Who are the major influencers when it comes to driving the conversation on social media with regard to healthcare ?

II Background

Social media has taken over most parts of modern life, and people use it extensively to help them make decisions. This is especially true when it comes to social media used for the purposes of healthcare. The public is able to gather information and make decisions in the form of recommendations that they receive from friends and family, opinion sharing through the internet in different forms and also by gathering information about the healthcare products and services that is available online. It is obvious that when people consume these products and services that they will have opinions about them, through which a lot of additional information about their intent can be understood. Of late, the percentage of people using social media to decide their healthcare choices has only expanded. A Pew study as of late revealed that 61% of American grown-ups look for health data on the web and 37% have gotten to client created health data on the web. Forty-two percent of all grown-ups say that they or somebody they know has been aided by taking after medical counsel or health data found on the web, a 43% expansion since 2006; just 3% of all grown-ups report that they or somebody they know has been hurt [3] .

III Related work

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.

IV Proposed Approaches

Our early approach will focus on :

1. Understanding the major keywords used to talk about healthcare 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 diseases 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 what kind of diseases. This would also include mapping the sentiment to understand how bad the degree of damage is for a disease in a region. 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.

Data Source: For this exploratory research, we will use Twitter data. Twitter allows users to post multiple features for a single tweet, including hashtags, images, retweeting, favoriting etc. This would allow for us to understand the intent behind the message of the users. We plan to gather close to 100,000 tweets which we believe will provide us with a large enough corpus to derive statistically significant results.

V System Architecture

Key components :
Data Mining : For this project, we will be gathering tweets about the healthcare industry from Twitter. We will be using relevant hashtags to gather information regarding the kind of tweets that will have to be mined. This can be done using twitter’s REST and Streaming API. We will be using the streaming API to gather tweets by using a persistent HTTPS connection.

Data Cleaning : Data cleaning comprises of the largest time chunk in a project. We will be using R (specially, the packages plyr, dplyr) to clean the data to extract useful keywords. The major intention os to model the data from semi structured to fully structured data.

EDA : Exploratory data analysis is important for this project as refining our choice of highlight factors that will be utilized later for examination is important. When we increase itemized nature with the information, we can return to the component choice stride since now as a result of EDA we may discover the elements we chose don't fill their expected need yet above all we may find different elements that add to the general picture the information presents.

Network Modelling : Network models would allow us to delve deeper into understanding the hierarchical structure of the nature of the relationships between different Twitter users to better understand who the influencers are in this scenario.

Visualization : Visualization can be used to convey a lot if information in a very simple manner. We plan to use Tableau and D3.JS for visualizations.

VI 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.

APPENDIX 1 – Tasks

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

APPENDIX 2 – Schedule

This is a tentative schedule to keep a track on progress of the project. This schedule may change or shift as development continues.

Project Checkpoint: 03/31

Project Presentation: 05/05

Final Report submission deadline: 05/12