2021

Non-Communicable Diseases and Social Media: A Heart Disease Symptoms Application

Ashwin Kumar Thandapani Kumarasamy

Daniel Asamoah

Ramesh Sharda

Title Page
Non-Communicable Diseases and Social Media: A Heart Disease Symptoms Application

Ashwin Kumar Thandapani Kumarasamy
Amazon.com, Inc.,
Seattle, WA, 98109 USA

*Daniel Adomako Asamoah
Wright State University,
Information Systems and Supply Chain Management Department,
Raj Soin College of Business,
Dayton, OH, 45435 USA
Email: daniel.asamoah@wright.edu

Ramesh Sharda
Oklahoma State University,
Management Science and Information Systems Department,
William Spears School of Business,
Stillwater, OK, 74078 USA

* Corresponding author

Keywords

Data analytics, social media, disease surveillance, epidemiology, non-communicable diseases

Non-Communicable Diseases and Social Media: A Heart Disease Symptoms Application

Abstract

Social media platforms have become ubiquitous and allow users to share information in real-time. Our study uses data analytics as an approach to explore non-communicable diseases on social media platforms and to identify trends and patterns of related disease symptoms. Exploring epidemiological patterns of non-communicable diseases is vital given that they have become prevalent in low income communities, accounting for about 38 million deaths worldwide.

We collected data related to multiple disease conditions from the Twitter microblogging platform and zoomed into symptoms related to heart diseases. As part of our analyses, we focused on the mechanism and trends of disease occurrences.

Our results show that specific symptoms may be attributed to multiple disease conditions and it is viable to identify trends and patterns of their occurrences using a structured analytics approach. This can then act as a supplementary tool to support epidemiological initiatives that monitor non-communicable diseases. Based on the study's results we identify that non-communicable disease surveillance approach using social media analytics could support the design of effective health intervention strategies.

Keywords

Data analytics, social media, disease surveillance, epidemiology, non-communicable diseases

Non-Communicable Diseases and Social Media: A Heart Disease Symptoms Application

1 Introduction

The amount of data produced every two days is more than what had been produced in all of human history prior to 2003 [1]. Social media platforms constitute a significant source of human generated data. For instance, on Twitter, users send about 31,250,000 messages and watch about 2,770,000 videos every minute on average [2]. Given the amount of data disseminated on social media platforms, there is tremendous potential to extract knowledge for practical applications such as disease surveillance.

Disease surveillance mostly involves a systematic process of collecting, analyzing and interpreting results from a large amount of health-related data. Results from the interpretation of such data can be used for the control, prevention, monitoring, resource planning and identification of risk factors regarding a disease condition. Whereas traditional clinical databases and methods are helpful for disease surveillance [3], information technologies such as social media and cellular phones are paramount to monitoring patterns of disease occurrence in large populations [4]. Such an effort can serve as a complementary tool to public healthcare initiatives to manage disease spread [5].

Knowledge extracted from social media platforms, through disease surveillance can facilitate timely intervention strategies [6]. Other benefits include accuracy and cost-effectiveness in tracking health sentiments, behaviors and outcomes that could support epidemiological decision making [7]. A disease's effect can either be sudden or slow depending on the propagation mechanism. For example the effects of the Ebola disease outbreak that hit several parts of Africa between 2014 and 2016 was immediate [8] whereas the lead poisoning effect in the Flint, Michigan drinking water problem took a relatively longer period to manifest [9]. Although multiple research studies have looked at identifying and tracking disease outbreaks with immediate effects such as the Ebola disease, surveillance research that targets non-communicable diseases such as the water crises in Flint, Michigan are uncommon.

To this end, we ask the question, can social media analytics be used for non-communicable diseases surveillance? We focus on non-communicable disease symptoms surveillance that develop at a slower pace and could culminate into a severe health outcome. This is an important study because unlike infectious disease epidemics which have far reaching consequences at

different geographic areas (such as the COVID-19 disease outbreak in 2019), non-communicable disease outbreaks have particularly high impact on low income communities since there are significantly fewer resources to monitor and combat them. Globally, non-communicable diseases are the leading causes of mortality, leading to about 38 million deaths yearly in low income communities [10]. Hence, whereas the impact of non-communicable diseases is not immediate and do not draw immediate and wide-spread media attention as infectious diseases do, their impact could cause long term health and financial burden on entire communities. Governments and health institutions, particularly in developing regions would hence benefit from timely identification of non-communicable diseases and subsequent design of intervention strategies. As one of the four most common non-communicable diseases alongside cancers, diabetes, and chronic respiratory diseases [11], we focused on heart diseases. Previous studies have shown that heart diseases are mostly preventable and caused by behavioral risk factors, unhealthy diet and lack of physical exercise [12]. Using the Apache Hadoop data analytics platform, we extracted and analyzed heart disease-related data from the Twitter microblogging platform. In our analyses, we determined if there is a trend in the occurrence of heart-related disease condition such in a way that should draw the attention of healthcare policy makers and institutions to a wider health problem.

In the next section, we discuss the literature related to disease surveillance and social media analytics' use to extract and analyze relevant data. We focus on the use of geo-location data for disease surveillance. Next, we discuss our methodology, including the big data platform, data collection, data preparation and management, symptoms categorization and noise reduction. We continue with the analysis, results, discussion, and implications. Although we classified the data into ten disease groups, we demonstrated our approach on only one group (with symptoms related to heart disease). We finally conclude with findings and suggestions based on our proof-of-concept approach.

2 Literature Review

2.1 Disease Surveillance and Non-Communicable Diseases

Disease surveillance is commonly described as a systematic activity consisting of collection, analysis, and interpretation of a large amount of health-related data. An overwhelming majority of disease surveillance initiatives and research focuses on communicable diseases. While surges in

some disease conditions could be epidemic with an immediate and negative impact in terms of lives lost, there may be upward trends in some non-communicable disease conditions that may not be an instantaneous, yet fatal to human health and livelihood in the long term. Such related diseases such as cancer, heart disease, and diabetes could be subtle, responding to minute changes in external factors and conditions [13]. For example, the Flint, Michigan health condition created by months of water poisoning resulted in fatal health conditions [9]. Whereas the water poisoning did not create a sudden epidemic, it was slow, yet steady in its impact as it took years to manifest. Prior to official channels of reporting related symptoms, could self-reported incidences on social media have helped health authorities recognize this potential threat and flagged it for further review?

2.2 Social Media and Disease Surveillance

Disease surveillance requires large data volumes from multiple sources. Bean et al. [14] explain what data should be reported to the Centers for Disease Control (CDC) during food-borne illness outbreaks. Likewise, Paquet et al. [15] explain in detail about the standards members in the European Union should adhere to in reporting diseases incidents to the European Union Center for Disease Prevention and Control (EUCDC). However, data reported to governmental agencies after symptoms have been reported may be too little and too late in preventing a health catastrophe. In contrast, self-reported data from social media platforms is available instantaneously and can be useful as an effective complementary tool for disease surveillance [16]. For instance, Brownstein et al. [17] explain in detail how social media data can be effective in identifying several disease outbreaks such as Salmonella and H1N1.

Social media applications for disease surveillance can be categorized into three main classes; epidemiologic monitoring and surveillance, situational awareness during emergency response and communication surveillance [5]. Whereas situation awareness and communication surveillance applications seek to predict needs (e.g. water supply) and disseminated public health information reception, epidemiologic monitoring and surveillance seek to predict disease incidence. Disease outbreak predictions are based on both official (CDC reports) and non-official (keyword-generated data from social media) data sources. As part of epidemiologic monitoring and surveillance, disease detection via syndromic surveillance focuses on self-reported data disclosed by individuals

[5]. We present our study as a syndromic surveillance to monitor disease conditions occurrences as reported on the Twitter microblogging platform.

Data from social media platforms such as Twitter has been shown to be helpful for building health application such as disease surveillance. Many previous health-related studies have utilized geo-tagged social media data such as from Twitter to generate insights for disease surveillance [18]. For instance, Broniatowski et al. [19] proposed an influenza surveillance system that is focused on geographic granularity. Their proposed model produced a 93% correlation between the data it generates and the data from CDC. While geo-tagged data may be extracted by either human readers or automated computing systems, structured processes for exploring and analyzing the large amount of data for monitoring public health trends still need to be developed and tested [7]. Big data applications and its use for public health-related monitoring and surveillance is non-trivial and presents significant technical challenges such as noise reduction [20].

Several studies have solely focused on non-communicable disease surveillance [21–24]. None of the prior studies have focused on leveraging social media analytics as a sole or supplementary tool for non-communicable disease surveillance. Social media platforms could be a promising approach to support non-communicable disease surveillance in low income communities due to the low cost of data access and faster development of healthcare-related applications. To this end, we believe our paper makes a significant contribution to the growing field of disease surveillance by designing an exploratory study to study how social media analytics can support this effort.

2.3 Disease Surveillance Methods

Several methods including statistical models [25], have been implemented in tracking and analyzing diseases epidemics. For instance, Aramaki et al. [26] proposed a support vector machine (SVM) based model to extract tweets related to influenza and analyze their propagation. The authors observed an 89% correlation between their proposed approach and the "*gold standard*" (Infection Disease Surveillance Center (IDSC) in Japan). The proposed SVM-based approach was better than the "*gold standard*" approach. A major drawback of the SVM approach, however, was constructing a training dataset manually (crowdsourcing) to train the SVM to differentiate between tweets that actually talk about illness and tweets that are irrelevant to the analysis. Likewise, Paul and Dredze [27] proposed an Ailment Topic Aspect Model Plus (ATAM+) to analyze tweets propagation related to several illnesses which belongs to the following categories; allergies,

depression, aches/pains, cancer, obesity, flu and dental. The ATAM+ model uses external resources to identify ailment categories [27]. Even though the ATAM+ can be used to model different disease categories, its performance is limited by the annotation process that is done manually [27]. Other limitations of some of the approaches relates to the accurate processing of informal syntax and spelling mistakes that are germane to the Twitter platform.

Given complex system requirements for non-communicable disease surveillance, previous research calls for multiple surveillance approaches to collecting and analyzing trends in non-communicable diseases [11]. Previous studies have also called for exploring and demonstrating a structured occurrence mechanisms of non-communicable diseases using data from social media platforms [28]. We extend the literature on public disease surveillance by exploring how social media analytics can be helpful in identifying trends in non-communicable diseases. Secondly, we respond to an identified gap in prior literature by laying the foundation for a structured social media analytics approach for analyzing non-communicable disease trends via social media platforms [7].

3 Methodology

3.1 Setting Up a Big Data Platform

3.1.1 Apache Hadoop

For this study, we used the Apache Hadoop Big Data platform [29] for data collection and analysis. The Apache Hadoop architecture is an open source infrastructure and one of the leading Big Data platforms used in both academia and industry.

Hadoop distributes huge amounts of data and computational processes to a large number of low-cost commodity machines. It consists of two key components namely, Hadoop Distributed File System (HDFS) (Fig. 1) and MapReduce [30]. MapReduce provides Hadoop with capabilities such as parallel computational power [31]. HDFS is a distributed file system, which is designed to function on commodity hardware [30]. HDFS is also a highly fault-tolerant file system as compared with traditional distributed file systems. It provides high throughput access to applications that access very large datasets [30].

Figure 1 here

Fig. 1 Hadoop file system architecture [30]

A typical Hadoop cluster comprises a single name node and multiple data nodes. All data on a Hadoop cluster are stored in the data nodes. The same data block is stored on multiple data nodes which enhances data availability and reliability as the data block can be accessible from other nodes even if a node containing a copy fails [29]. The name node keeps track of which data blocks of which files are stored on individual data nodes.

We used HDFS as the underlying subproject for storing tweets. We also used MapReduce jobs to access and analyze tweets on the HDFS. The Hadoop cluster we designed and configured consisted of a single master node and 6 data nodes. We streamed data from Twitter API to HDFS using Apache Flume. We also utilized TweetNLP [32] and Google Fusion Tables [33]; the former for noise reduction and the latter for geo-location enabled data visualization.

The Apache Hadoop platform includes other sub-projects such as Apache Flume and Apache Hive [29]. The Hadoop subprojects and other tools we used for data collection and analysis are described in subsequent sections.

3.1.2 Apache Flume

Apache flume is an open source software used for collecting, grouping and moving large quantities of data [34]. Apache flume can be used to transport data related to social media, logs, and network traffic. In this study, we used Apache Flume to stream social media data from Twitter to the HDFS. Important components of a flume agent are source, channel and sink as shown in Fig. 2. A Flume source consumes events handed to it by an external entity such as a web server. A channel stores events until a sink consumes them. A sink puts the events from the channel into an external storage.

Figure 2 here

Fig. 2 Data flow architecture - Apache flume [34]

3.1.3 Apache Hive

Apache Hive provides an SQL-like interface to a large dataset that is stored in the HDFS [29]. Hive also allows the user to impose a structure onto the data and query the data using SQL-like language called HiveQL [29]. In this study, we used Apache Hive to impose structure on data from Twitter and prepare it for further data analysis.

3.1.4 Noise and TweetNLP

As is consistent with prior research, we collected tweets using diseases-related keywords [35]. However, some keywords used in data collection can be used in many other contexts beside conversation on a disease condition. Tweets with such keywords introduce noise and render some collected tweets useless for analysis. This is an impediment to social media's use for health surveillance [7]. Hence, we used a Natural Language Processing Tool (NLP) called TweetNLP [32] to identify the occurring keyword's context and to decide whether the context was useful for the analysis. TweetNLP, developed by researchers at Carnegie Mellon University is a java-based tokenizer and a parts-of-speech tagger for Twitter data.

3.1.5 Google Fusion Table

Google Fusion tables aid in collecting, visualizing and sharing datasets over the web [33]. Data visualization in a Google Fusion table involves creating charts, maps, graphs, and custom layouts in a short period. We used Google Fusion table to visualize geo-location enabled tweets. This is intuitive as it provides insights about disease patterns in a given geographic region in a given time interval.

3.2 Data Collection

To collect data from Twitter, we created a Twitter application which enabled us to acquire multiple authentication codes including a Consumer Secret, Consumer Token, Access Token and Access Token Secret from Twitter. We also created a custom flume agent that would authenticate and collect data from the Twitter API based on a set of keywords. Tweets were received in JSON format and were stored in HDFS. All these authentication parameters, keywords and storage path were located on the HDFS and were stored in a flume configuration file.

Extraction and use of data from social media platforms could present a host of challenges to healthcare applications. Islam et al. [28] outlines some of these challenges but as a research tool the key obstacles include lack of evidence, low-quality information, participants confidentiality, commercial interests by industry players, digital divide and participants preferences in using different social media platforms.

Since it is a developing research area, strategies to combat the outlined challenges are still evolving. In our study, we reduced noise by targeting and focusing on a specific topic which is

heart conditions related to non-communicable diseases. Also, the TweetNLP tool that we used helped to ensure that quality information was generated. Furthermore, in order to generate relevant analysis, the researchers trained in the architecture, use and, propagation of data on the Twitter microblogging platform.

Multiple approaches, including methods that use disparate systems such as Apache Hadoop and the Amazon Mechanical Turk platform have been used for topic-specific data collection and management [36]. We used a parsimonious method where we first identified keywords that were related to ten disease conditions namely "*heart disease, stroke, diabetes, flu, STD, diarrhea, tiredness, muscle spasm, liver disease, and cancer*". Tweets, which were relevant to our analysis were segregated using TweetNLP [32]. The keys words were generated based on a health-related data repository hosted by two main sources. The first repository is Medline Plus, a service provided by the U.S. National Library of Medicine through the National Institute of Health [37]. The second, MedicineNet, is a popular database that hosts medical data on various disease conditions [38]. We collected health-related tweets for two weeks. This amounted to about 10GB of tweets. A sample of the keywords used are listed below:

> *Agina, Shortness of breath, irregular heartbeats, high blood pressure, stents, angioplasty, bypass surgery, pace maker, ace inhibitor, aspirin, beta-blockers, cough, rash, high potassium, numbness of face, numbness of arms, numbness of legs, loss of vision, loss of coordination, loss of speech, dizziness, ECG, EKG, clot dissolving medicine, aspirin, anti-platelet, loss of abilities, loss of motor skills, loss of speech, increased urination, fatigue, high blood glucose levels, insulin, loss of vision, kidney damage, nerve damage, fever, aches and pains, cough, runny nose, decongestant, immune stimulation, hyperthermia, vaccine, antitoxin, thymus extract, sono-photo dynamic, detoxification, cleansing, enema, chemotherapy, radiation, immune suppression, typhlitis, gastrointestinal distress, nausea, vomiting, anorexia, diarrhea, cramps*

Whereas the data extraction process is based on keyword matching, we do not propose to wholly infer specific diseases based only on keywords. A conclusive determination of a disease's gradual manifestation would require support from medical diagnosis and other scientific experiments. We present our approach as only an auxiliary means for non-communicable disease surveillance.

3.3 Data Preparation and Management

The tweets collected from Twitter were in JSON format and included variables such as username, timestamp for tweet, location data, re-tweet count, tweet language and time zone. Since the JSON data was semi-structured, we created a schema for it using Apache Hive. Apache Hive allowed us to structure our data like a traditional database and query the data using HiveQL. We initially started by creating a table for our tweets with columns including id, created at timestamp, tweet's language, re-tweet count, text, and user. The HiveQL script for creating the table is shown in Fig. 8 in Appendix A. After the data was structured, we exported it data back to HDFS for analysis.

3.4 Grouping Symptoms into Relevant Categories

In order to perform further analysis on the tweets, we grouped keywords used to retrieve the Twitter data into categories that correspond to specific disease conditions as shown in Table 1. This resulted in 10 groups representing disease conditions such as heart attack, stroke, kidney stones, cancer, and asthma. There were some symptoms that were present in multiple groups.

After grouping the symptoms, we created a clustering algorithm that read every tweet collected for the symptoms listed above and placed the same in a corresponding group. This algorithm eliminated any tweet without keyword from one of the groups.

Table 1 here

Table 1 Symptoms and their corresponding disease groups

In the algorithm, the *Main()* method is the main entry point for the clustering algorithm and gets executed first whenever the clustering algorithm is called. This method calls a set of functions necessary for the data preparation process. The functions are explained in Appendix B. Noise elimination and clustering occur simultaneously. A detailed noise eliminator process description is shown in the clustering algorithm pseudo-code in Fig. 9 in Appendix A.

3.5 Noise Reduction Using NLP Tool and Parts-of-Speech Tagging

There was considerable amount of noise in the collected tweets. This usually occurs when some keywords in the collected tweets are out of context and irrelevant to the conversation of interest. To aid in eliminating noise and removing irrelevant tweets, we developed a JAVA application that made use of TweetNLP to eliminate noise and used it along with the clustering algorithm implementation. It identified the keywords occurrences' context and aided in determining whether

a tweet was useful for the analyses or not. The noise eliminator helped remove tweets that were not relevant to our discussion's subject. The noise eliminator's functions are *analyzePOSTaggerResults()* and *isNoise()* and are described in Appendix C. We leveraged POS for our analysis since it has been shown to contribute to tweet objectivity [39].

4 Analysis and Results

In this section, we focus on a single group as a demonstration of a structured process to analyze a large amount of social media data for non-communicable disease surveillance, and present corresponding results. Our analyses only focus on English tweets and support disease surveillance for only heart diseases (Group 1). We focused on heart diseases since it is one of the four most common non-communicable diseases alongside cancers, diabetes, and chronic respiratory diseases [11]. It is mainly caused by preventable behavioral risk factors, unhealthy food intake and lack of physical exercise [12]. Similar to other major non-communicable diseases, heart disease prevalence in developed countries remain constant while its prevalence in developing regions such as Africa and South-East Asia is increasing. This calls for more research in this area so that effective epidemiological strategies can be designed to combat its prevalence.

The procedure for this analysis in this study may be extended to other disease groups. We focus on geo-tagged tweets, which is a small percentage of the total tweets collected as shown in Fig. 3. This phenomenon is consistent with prior literature on geo-location social media analysis [7].

Figure 3 here

Fig. 3 Distribution of tweets collected and cleaned

The keywords distribution for Group 1 over two weeks is shown in Fig. 4. Keywords such as rash, high potassium and high blood pressure were the most occurring terms related to heart diseases as is shown in Group 1. This is consistent with prior medical studies that have indicated that skin rashes in unusual spots is a heart disease symptom for heart infection [40].

Figure 4 here

Fig. 4 Distribution of keywords for group 1 (Heart diseases)

Fig. 4 indicates that rash was the most prevalent disease symptom as observed in our data set. Users who posted these symptoms could be suffering from any of the three types of heart

inflammations; pericarditis, myocarditis or endocarditis [41]. Besides disease condition, the analyses show associated treatment methods. For instance, Fig. 4 shows that one of the most common diagnostics tools used for heart diseases was the electrocardiography (ECG). Again, this is consistent with treatment approaches in practice [42].

Messages on diseases and various conditions were posted at varying frequencies during different times in the day. Monitoring these keywords periodically could provide in-depth insight into how symptoms progress over a period. We identified what the prominent symptoms leading to heart diseases were. Fig. 5 depicts a keyword distribution on an hourly schedule based on the central time zone. For instance, we could determine if a symptom such as rash is reported consistently throughout the day or it was common during certain times of the day. Fig. 5 also shows that more symptoms were reported during early morning and late-night periods than periods after midnight and mid-day.

As a disease surveillance mechanism, understanding the relationship between the different keywords' occurrences was beneficial in understanding the disease's propagation trends and patterns. For instance, Fig. 6 shows the relationship between high blood pressure and ECG. The regression lines provide an intuitive insight about the interaction between these variables. High blood pressure does not typically portray outward symptoms. An ECG test is, therefore, useful in determining if an individual has high blood pressure. The trend as shown by the regression lines in Fig. 6 indicates that as high blood pressure incidence decreased, ECG mentions, and by extension, its use decreased. Typically, in such analysis, the results should be interpreted within the right context and in consideration of other keywords that are involved. For instance, the reason for the dip in ECG on November 26th, when there was an increase in high blood pressure, may be attributed to the relationship of these keywords with keywords in other groups.

Figure 5 here

Fig. 5 Distribution of keywords for group 1 (Heart diseases) for 1 day

The disease symptoms' geo-locations were plotted on a map as shown in Fig. 7. It shows a distribution of related symptoms for Group 1 in different geographic areas. Whereas such outputs are limited by the number of self-reported incidences around the globe, it may be vital in supporting other epidemiolocal tools and processes to identify disease trends.

Figure 6 here

Fig. 6 Analysis of linear relationship between high blood pressure and ECG

Figure 7 here

Fig. 7 Plot of geo-location enabled tweets belonging to group 1

5 Discussion and Implications

Using social media and mobile network capabilities for public health surveillance is common for infectious disease epidemics [6]. Previous studies have targeted epidemics and pandemics such as COVID-19 [43] and the 2014 Ebola outbreak which recorded over 28,000 suspected, probable and confirmed cases [44]. Although disease epidemics could be devasting to almost all regions, non-communicable diseases are particularly devastating to low- and middle-income regions. For instance, during non-communicable diseases manifestations, up to 80% of related deaths do occur in low and middle income countries [28]. Besides, most symptoms of non-communicable diseases occur and propagate without much notice. In this study, we explored the use of Twitter microblogging platform for non-communicable disease surveillance. We leveraged social media analytics to extract, clean, and analyze the data.

Our study included a time-based analysis of publicly available health information on the Twitter microblogging platform. We extracted geo-location data from tweets and analyzed disease patterns, focusing on heart disease occurrences in different parts of the world. Fig. 7 shows a plot of the geo-location data extracted from a Google Fusion Table [33]. Geo-location data is vital to disease surveillance as it provides key insights such as the outbreak's source, propagation speed and number of people affected per a particular area. Such an analysis offers healthcare stakeholders the ability to make granular epidemiology decisions [45].

Our study contributes to public healthcare management as it could support vital intervention strategy design for non-communicable yet, catastrophic diseases. When geo-location data is extracted from social media, it can help decision makers understand the disease propagation from across geographic regions. Such decision tools could help determine which location and at what point to introduce relevant intervention resources. These resources could include personnel, equipment, and other logistics. For instance, when there is an increasing number of posts on a set of related symptoms in a specific geographic region, this might trigger close monitoring of

ailments reported in that region. Stakeholders such as epidemiologists could then perform a follow-up study of the relationships between the observed symptoms on social media platforms and the reported ailments.

The use of social media platforms for disease surveillance is important given that such platforms have become a primary information source by an increasing number of individuals [46]. Individuals may not readily respond to a traditional survey about a disease outbreak but would easily seek and provide information on a social media platform about a disease condition. Disease patterns that are identified via social media platforms may provide preliminary insights into imminent disease outbreaks and allow adequate intervention strategies to be designed to handle related occurrences that may emerge.

Also, whereas traditional means of disease surveillance provides a retrospective disease epidemiological perspective, social media use could support cost-effective disease occurrence predictions. We propose that the two approaches could co-exist such that health institutions and healthcare policy stakeholders could utilize the output from social media-based disease surveillance to support tradition disease surveillance practices. Our non-communicable disease surveillance approach is consistent with expressed need for testing and developing structured procedures for analytics-based disease surveillance procedures on social media [7].

In the past few years, data management capabilities have paved the way for advanced big data applications in healthcare where several epidemiological models have been created for infectious disease surveillance [45]. Data-driven epidemiological models for non-communicable diseases, and best practices on how such models can be developed and executed using social media platforms is lacking. Our study provides a proof of the concept of using social media data for non-communicable disease surveillance. This offers an optimistic data-oriented complementary approach to healthcare institutions and governments as they battle the stealth spread of non-communicable diseases.

Finally, we contribute to the methodology literature as we use the Apache Hadoop big data analytics platform to extract, manage, and analyze health-related social media data. Whereas social media data alone may not be enough for non-communicable disease surveillance, such data can support efforts by healthcare practitioners in intervention strategies development.

6 Conclusion

In this study, we explored how social media could be helpful for non-communicable disease epidemiology. Using big data analytics technology, we also showed how analyzing the symptoms distribution of a particular disease can help monitor the disease's propagation and identify if its occurrence and trends is significant enough to trigger the attention of policy makers and healthcare institutions. Geo-location data inclusion helped provide valuable insights such as disease symptom's propagation patterns. Preliminary results presented show the potential of social media platforms, a geo-located based source of data, for disease surveillance. Insights based on our study could help healthcare practitioners to better track and manage non-communicable disease outbreaks. Intervention strategies can also be implemented in a timely manner and at the right locations.

As a potential study limitation, there is a relatively small number of geo-tagged tweets as compared to the total number of tweets in the data set. This is because not all users typically provide accurate location data in their profiles. While the absence of enough geo-tagged tweets is consistent with previous studies [7], follow-up studies would include secondary geographical data so as to enrich our location-based analysis.

Also, for a specific study, the duration of tweet extraction could be a factor in ensuring that the corpus is representative of the population of interest. In our study, we collected tweets over a period of two weeks. We recognize that a longer period would be ideal for a non-communicable disease which manifests at a slower pace.

We focused on only English tweets. However, the processes for disease surveillance could be extended to other languages and specific geographic regions. This would help healthcare stakeholders in extracting more granular and relevant insights. Future studies can also use established research approaches such as design science and case studies to explore specific non-communicable disease propagation phenomena in different geographic regions.

Lastly, given that social media analytics for healthcare applications is a developing field, issues related to data privacy and quality would need recurring attention. This is only an exploratory study and in future, more advanced data extraction strategies beyond keyword-based analysis and advanced natural language processing methods should be explored.

References

1. Bibri SE. Data Science for Urban Sustainability: Data Mining and Data-Analytic Thinking in the Next Wave of City Analytics. Urban B Ser. 2018.
2. Marr B. Big Data: 20 Mind-Boggling Facts Everyone Must Read [Internet]. Forbes Enterp. Cloud. 2019 [cited 2019 May 11]. Available from: https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#3080d8a717b1
3. Grais RF, Ellis JH, Kress A, Glass GE. Modeling the Spread of Annual Influenza Epidemics in the US: The Potential Role of Air Travel. Health Care Manag Sci. Springer; 2004;7:127–34.
4. Ouedraogo B, Gaudart J, Dufour JC. How does the Cellular Phone Help in Epidemiological Surveillance? A Review of the Scientific Literature. Informatics Heal. Soc. Care. 2019.
5. Fung ICH, Tse ZTH, Fu KW. The Use of Social Media in Public Health Surveillance. West Pacific Surveill Response J WPSAR. 2015;6.
6. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLoS One. 2015.
7. Mogo E. Social Media As A Public Health Surveillance Tool: Evidence And Prospects [Internet]. Baltimore; 2018. Available from: https://enterprise.sickweather.com/downloads/SW-SocialMedia_WhitePaper.pdf
8. Painter JE, von Fricken ME, Viana de O. Mesquita S, DiClemente RJ. Willingness to Pay for an Ebola Vaccine During the 2014–2016 Ebola Outbreak in West Africa: Results from a U.S. National Sample. Hum Vaccines Immunother. 2018;14:1665–71.
9. Morckel V. Why the Flint, Michigan, USA Water Crisis is an Urban Planning Failure. Cities. 2017;62:23–7.
10. Al-Mawali A. Non-communicable diseases: shining a light on cardiovascular disease, Oman's biggest killer. Oman Med J. Oman Medical Specialty Board; 2015;30:227.
11. Kroll M, Phalkey RK, Kraas F. Challenges to the surveillance of non-communicable diseases - A review of selected approaches. BMC Public Health [Internet]. BioMed Central Ltd.; 2015 [cited 2021 May 21];15:1–12. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-015-2570-z
12. WHO | Global status report on noncommunicable diseases 2010 [Internet]. [cited 2021 May 21]. Available from: https://www.who.int/nmh/publications/ncd_report2010/en/
13. Misganaw A, Mariam DH, Ali A, Araya T. Epidemiology of Major Non-Communicable Diseases in Ethiopia: A Systematic Review. J Health Popul Nutr. BioMed Central; 2014;32:1.
14. Bean NH, Goulding JS, Daniels MT, Angulo FJ. Surveillance for Foodborne Disease Outbreaks - United States, 1988-1992. J. Food Prot. 1997.
15. Paquet C, Coulombier D, Kaiser R, Ciotti M. Epidemic Intelligence: A New Framework for Strengthening Disease Surveillance in Europe. Euro Surveill. 2006.
16. Jordan SE, Hovet SE, Fung ICH, Liang H, Fu KW, Tse ZTH. Using twitter for public health surveillance from monitoring and prediction to public response [Internet]. Data. MDPI AG; 2019

[cited 2021 May 21]. p. 6. Available from: www.mdpi.com/journal/data
17. Brownstein JS, Freifeld CC, Madoff LC. Digital Disease Detection - Harnessing the Web for Public Health Surveillance. N. Engl. J. Med. 2009.
18. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Predicting Flu Trends Using Twitter Data. 2011 IEEE Conf Comput Commun Work INFOCOM WKSHPS 2011. 2011.
19. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. PLoS One. 2013;
20. Culotta A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. SOMA 2010 - Proc 1st Work Soc Media Anal. 2010.
21. Asgari F, Aghajani H, Haghazali M, Heidarian H. Non-Communicable Diseases Risk Factors Surveillance in Iran. Iran J Public Health [Internet]. 2009 [cited 2021 May 26];38:119–22. Available from: https://ijph.tums.ac.ir/index.php/ijph/article/view/2867
22. Alwan A, MacLean DR, Riley LM, D'Espaignet ET, Mathers CD, Stevens GA, et al. Monitoring and surveillance of chronic non-communicable diseases: Progress and capacity in high-burden countries [Internet]. Lancet. Elsevier B.V.; 2010 [cited 2021 May 26]. p. 1861–8. Available from: http://www.thelancet.com/article/S0140673610618533/fulltext
23. Morais AL, Rijo P, Batanero Hernán MB, Nicolai M. Biomolecules and Electrochemical Tools in Chronic Non-Communicable Disease Surveillance: A Systematic Review [Internet]. Biosensors. MDPI AG; 2020 [cited 2021 May 26]. Available from: https://pubmed.ncbi.nlm.nih.gov/32927739/
24. Kim Streatfield P, Khan WA, Bhuiya A, Hanifi SMA, Alam N, Bagagnan CH, et al. Adult non-communicable disease mortality in Africa and Asia: Evidence from INDEPTH Health and Demographic Surveillance System sites. Glob Health Action [Internet]. Co-Action Publishing; 2014 [cited 2021 May 26];7. Available from: https://pubmed.ncbi.nlm.nih.gov/25377326/
25. McAree PW, Bauer KW, Louis DJ, Jackson JA. Use of Statistical Process Control for Surveillance of Pulmonary Dysfunction in Groups in the Workplace. Health Care Manag Sci. Springer; 1998;1:53–9.
26. Aramaki E, Maskawa S, Morita M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. EMNLP 2011 - Conf Empir Methods Nat Lang Process Proc Conf. 2011.
27. Paul MJ, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. Int'L AAAI Conf Weblogs Soc Media. 2011.
28. Islam SMS, Tabassum R, Liu Y, Chen S, Redfern J, Kim S-Y, et al. The Role of Social Media in Preventing and Managing Non-Communicable Diseases in Low-and-Middle Income Countries: Hope or Hype? Heal Policy Technol. Elsevier; 2019;8:96–101.
29. The Apache Software Foundation. Welcome to Apache™ Hadoop [Internet]. 2019 [cited 2016 Dec 15]. Available from: http://hadoop.apache.org/
30. Borthakur D. HDFS architecture guide. Hadoop Apache Proj [Internet]. 2008;53. Available from: http://archive.cloudera.com/cdh/3/hadoop-0.20.2-cdh3u6/hdfs_design.pdf%5Cnpapers3://publication/uuid/BE03DF70-D0C1-441E-A65F-1888C84992D6

31. Dean J, Ghemawat S. MapReduce : Simplified Data Processing on Large Clusters. Commun ACM [Internet]. 2008;51:1–13. Available from: http://portal.acm.org/citation.cfm?id=1327492
32. Owoputi O, O'Connor B, Dyer C, Gimpel K, Schneider N. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. CMU-ML-12-107. 2012.
33. Gonzalez H, Halevy AY, Jensen CS, Langen A, Madhavan J, Shapley R, et al. Google Fusion Tables : Web-Centered Data Management and Collaboration. Proc 2010 ACM SIGMOD Int Conf Manag data [Internet]. 2010. p. 1061–6. Available from: http://dl.acm.org/citation.cfm?id=1807286
34. The Apache Software Foundation. Apache Flume [Internet]. 2019. Available from: http://flume.apache.org/
35. Nagel AC, Tsou MH, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The Complex Relationship of Realspace Events and Messages in Cyberspace: Case Study of Influenza and Pertussis Using Tweets. J Med Internet Res. 2013;
36. Duberstein S, Doran D, Asamoah DA, Schiller S. Finding and Validating Medical Information Shared on Twitter: Experiences Using a Crowdsourcing Approach. Int J Web Eng Technol. 2019;14:80–98.
37. Medline Plus. Service of the US National Library of Medicine and the National Institutes of Health [Internet]. 2014. Available from: http://medlineplus.gov
38. MedicineNet. No Title [Internet]. 2014. Available from: http://www.medicinenet.com/script/main/hp.asp
39. Asamoah DA, Sharda R. What should I believe? Exploring information validity on social network platforms. J Bus Res. Elsevier; 2020;122:567–81.
40. Mayo Clinic. Heart Disease Symptoms [Internet]. Mayo Found. Med. Educ. Res. 2014. Available from: http://www.mayoclinic.org/diseases-conditions/heart-disease/basics/symptoms/con-20034056
41. Pawsat DE, Lee JY. Inflammatory disorders of the heart: pericarditis, myocarditis, and endocarditis. Emerg Med Clin North Am. Elsevier; 1998;16:665–81.
42. Ferrari E, Imbert A, Chevalier T, Mihoubi A, Morand P, Baudouy M. The ECG in pulmonary embolism: Predictive value of negative T waves in precordial leads - 80 Case reports. Chest. 1997;111:537–43.
43. Zheng YY, Ma YT, Zhang JY, Xie X. COVID-19 and The Cardiovascular System. Nat Rev Cardiol. 2020;17:259–260.
44. Mobula LM, Nakao JH, Walia S, Pendarvis J, Morris P, Townes D. A Humanitarian Response to the West African Ebola Virus Disease Outbreak. J Int Humanit Action. SpringerOpen; 2018;3:10.
45. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big Data for Infectious Disease Surveillance and Modeling. J Infect Dis. Oxford University Press; 2016;214:S375–9.
46. Pew Research Center. Social Media Fact Sheet [Internet]. 2019 [cited 2020 Apr 21]. Available from: https://www.pewresearch.org/internet/fact-sheet/social-media/

Appendix A: Script for Creating Tweet Table and Cluster Algorithm Pseudocode

Figure 8 here

Fig. 8 HiveQL script for creating the tweet table

Figure 9 here

Fig. 9 Pseudo-code for clustering algorithm

Appendix B: Functions in Main Method

clusterTweets(): This method is used to cluster tweets based on keywrods occurrences. The method also calls the findCluster() and cleanText() methods in a TweetClusterer class.

folderVerifier(): This method verifies whether the input path specified in HDFS is valid or not.

*getAllFilePAthsInAFolder():*This method is used to get all files stored under a folder in HDFS.

*getKeywordMatched():*This method is used to return any keyword that matches with the tweet text.

*writeToCSVFile():*This method writes clustered tweets onto local storage in CSV format.

Methods of the TweetClusterer class were used in data preparation and are defined as follows:

findCluster(): This method is used to find the appropriate cluster to which a tweet belongs.

cleanText(): This method calls isNoise() in Noise Eliminator class. It is used to determine whether the tweet is relevant to our analysis or not.

Appendix C: Functions in Noise Eliminator

analyzePOSTaggerResults(): This method calls TweetNLP and it returns a tweet with its corresponding parts of speech (POS) tags.

isNoise(): This method calls *analyzePOSTaggerResults()* method and distinguishes noise from tweets useful for our analysis based on the POS tags obtained from TweetNLP.

Figures

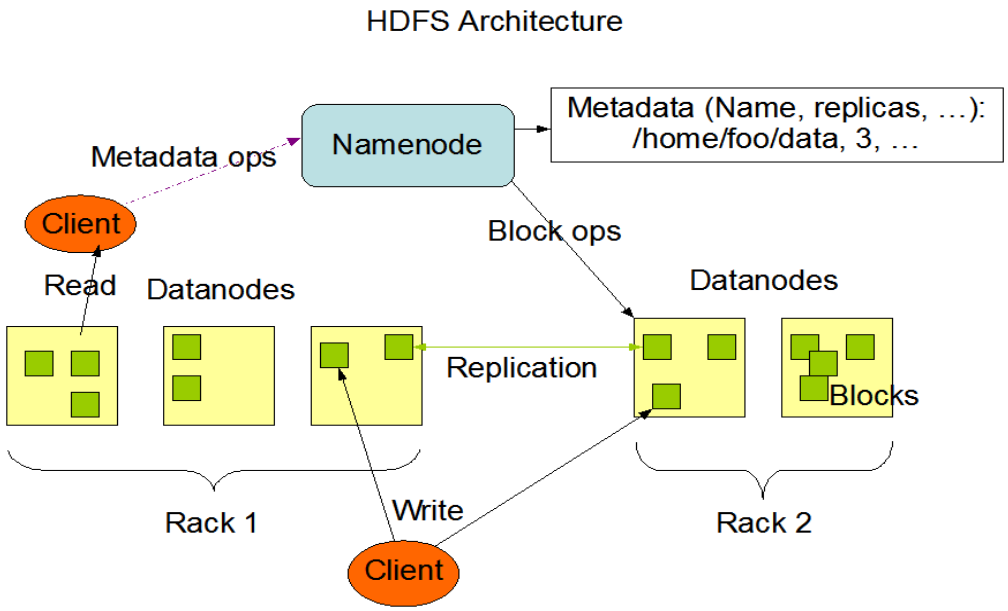

Fig. 3 Hadoop file system architecture [30]

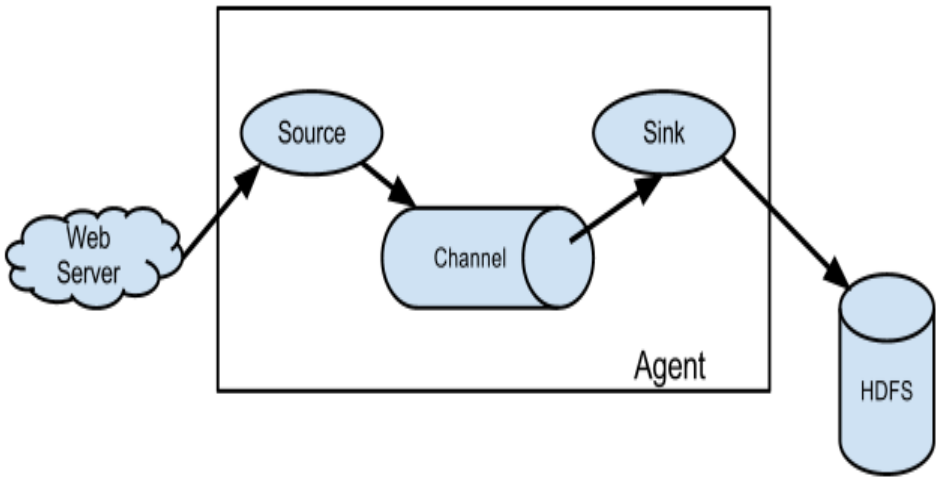

Fig. 4 Data flow architecture - Apache flume [34]

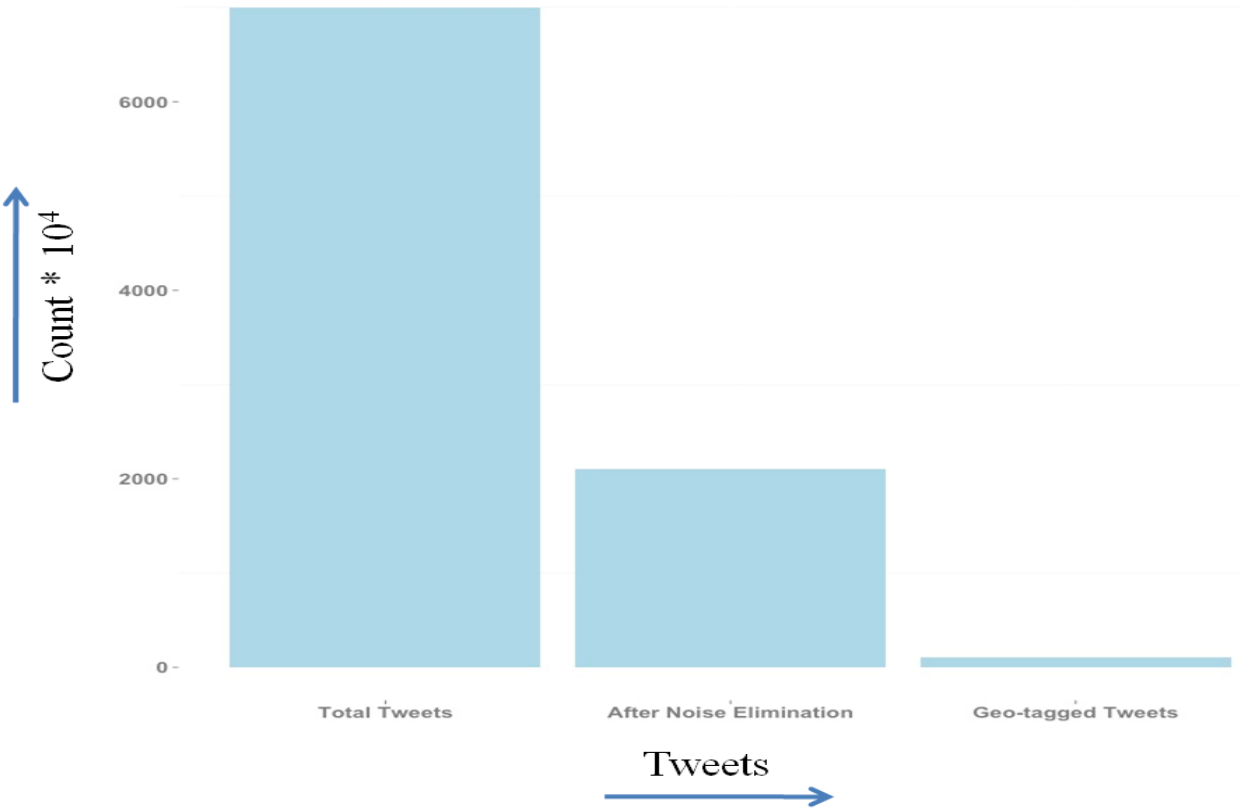

Fig. 3 Distribution of tweets collected and cleaned

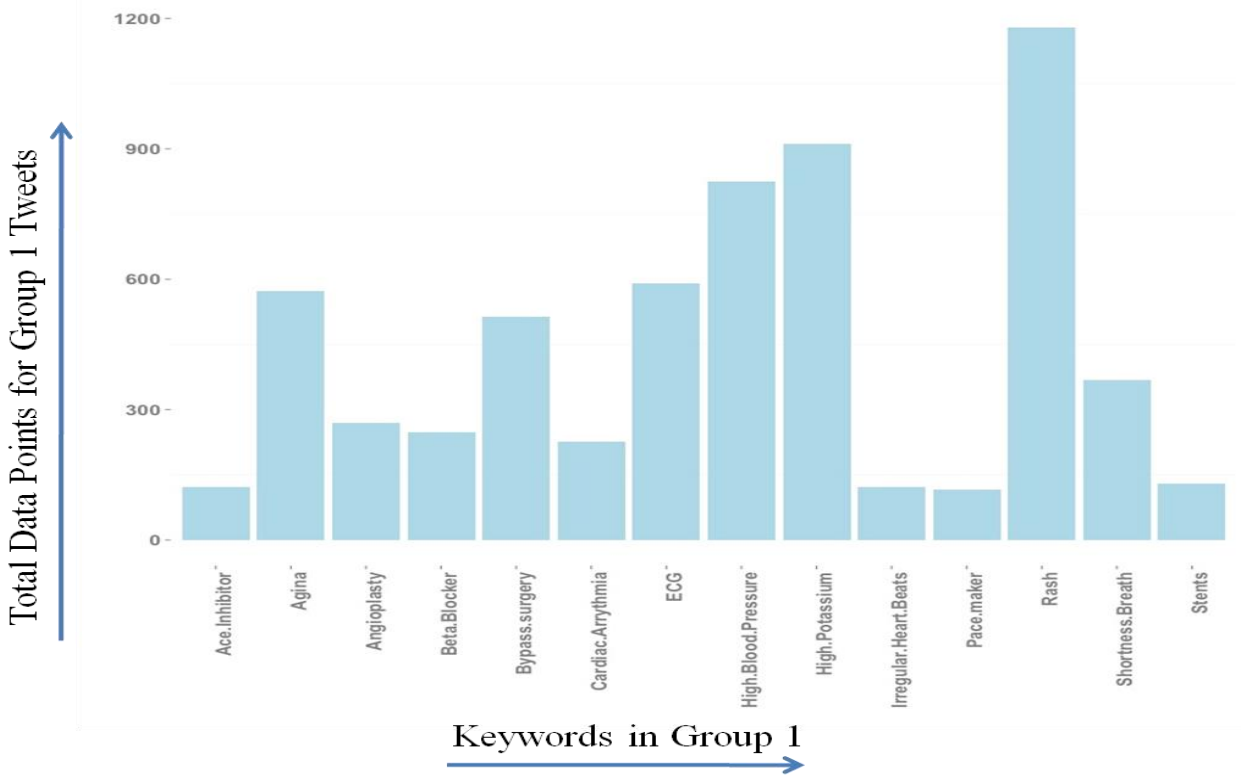

Fig. 4 Distribution of keywords for group 1 (Heart diseases)

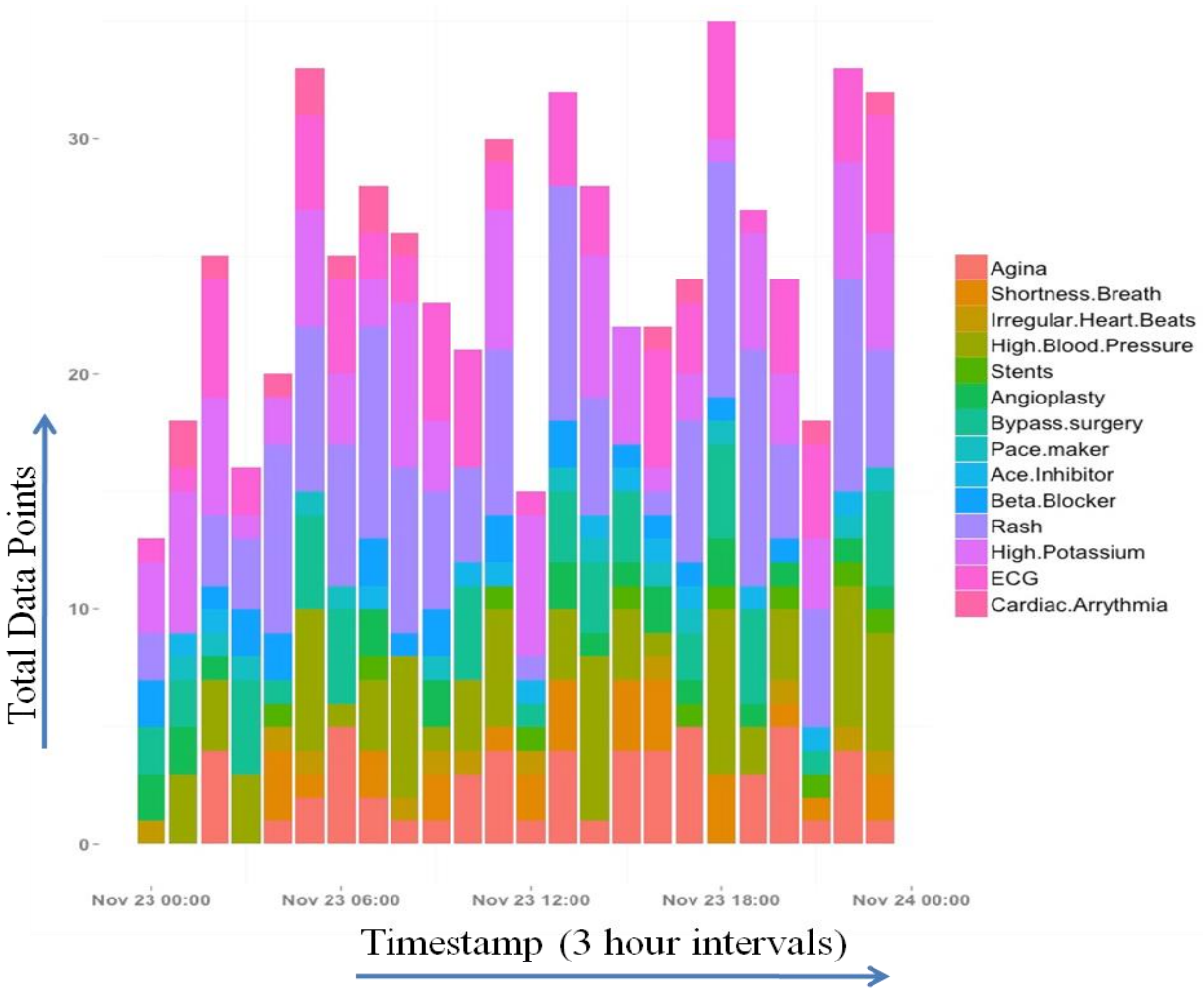

Fig. 5 Distribution of keywords for group 1 (Heart diseases) for 1 day

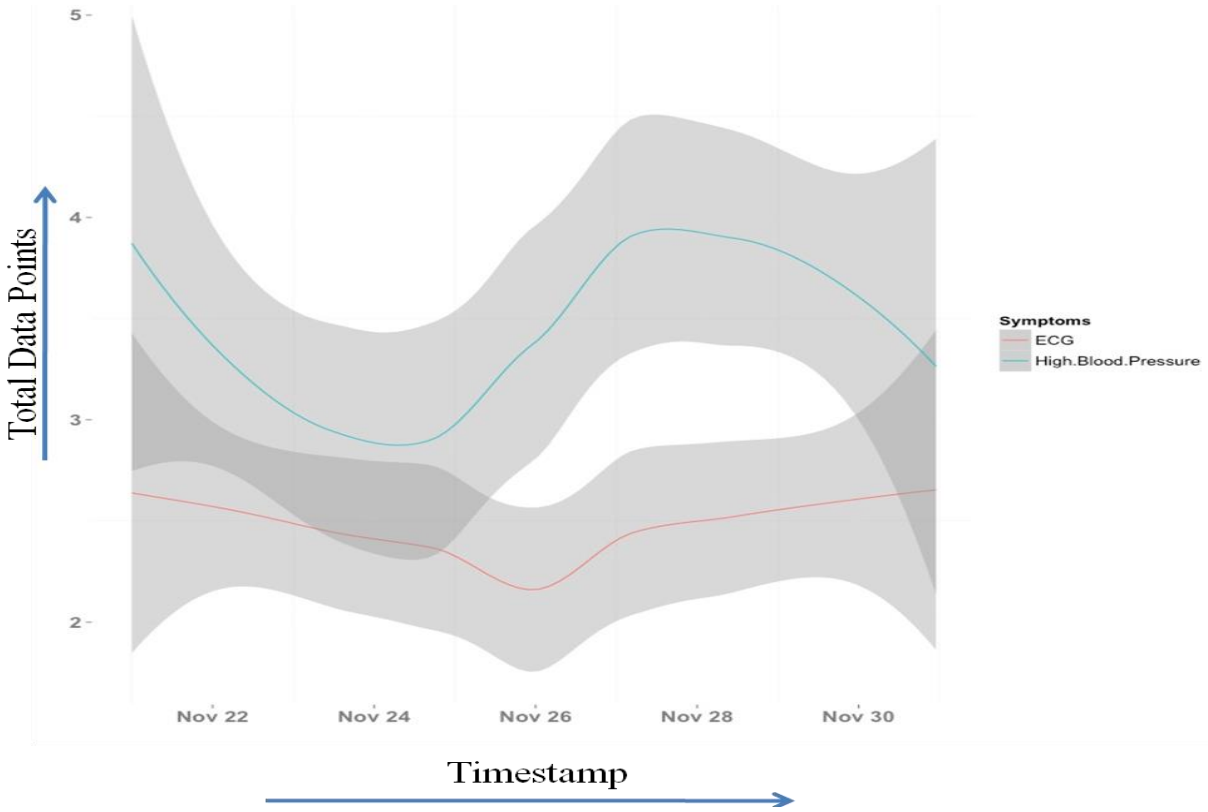

Fig. 6 Analysis of linear relationship between high blood pressure and ECG

Fig. 7 Plot of geo-location enabled tweets belonging to group 1

```
CREATE EXTERNAL TABLE Tweets (
id BIGINT,
created_at STRING,
lang STRING,
retweet_count INT,
retweeted_status STRUCT< id: BIGINT, created_at: STRING, entities: STRUCT<
hashtags:ARRAY<STRUCT<text:STRING>>>, text:STRING, lang:STRING, user:
STRUCT<screen_name:STRING,name:STRING, `location`: STRING,statuses_count:INT,
time_zone:STRING>, retweet_count:INT>,
entities STRUCT<hashtags:ARRAY<STRUCT<text:STRING>>>,
text STRING,
user STRUCT< screen_name:STRING, name:STRING, statuses_count:INT, `location`:
STRING, time_zone:STRING>,
in_reply_to_screen_name STRING)
ROW FORMAT SERDE 'com.hive.serde.JSONSerDe' LOCATION
'/user/hadoopusr/flume_data/tweets';
```

Fig. 8 HiveQL script for creating the tweet table

```
clusterTweets (tweet){

    1. Clean the tweet.
       cleaned_tweet = cleantext(tweet);
    2. Find the cluster which the tweet belongs.
       Cluster = findCluster(cleaned_tweet);
    3. If the tweet doesn't belong to any cluster terminate the
       function
       If (cluster==null)
           Return;
       End if
    4. Find whether the tweet is noise. If the tweet is noise return
       nothing else return the cluster which the tweet belongs.
       is_tweet_noise = eliminateNoise(cleaned_tweet,cluster);
       If (is_tweet_noise) then
           Return;
       Else
           Return cluster;
       End if
}
```

Fig. 9 Pseudo-code for clustering algorithm

Non-Communicable Diseases and Social Media: A Heart Disease Symptoms Application

Tables

Table 2 Symptoms and their corresponding disease groups

Group number	Disease condition	Corresponding symptoms
1	Heart Disease	Agina, Shortness breath, irregular heartbeats, high blood pressure, stents, angioplasty, bypass surgery, pace maker, ace inhibitor, beta-blockers, rash, high potassium, ECG, cardiac arrythmia
2	Stroke	Numbness of face, numbness of arms, numbness of legs, loss of vision, loss of coordination, loss of speech, dizziness, EKG, ECG, clot dissolving medicine, aspirin, anti-platelet, loss of abilities, loss of motor skills, loss of speech
3	Diabetes	Increased urination, fatigue, blood glucose levels, insulin, loss of vision, kidney damage, nerve damage
4	Flu	Aches pains, cold, cough, fever, runny nose, decongestant, cough syrup, cough tablet, cough tablets
5	STD	Shingles, herpes
6	Diarrhea	Diarrhea, constipation, detoxification stomach, detoxification gut, detoxification body, cleansing stomach, cleansing gut, enema, gastrointestinal distress, typhlitis, nausea, vomiting
7	Tiredness	Anemia, fatigue
8	Muscle Spasm	Cramps
9	Liver Disease	Hepatitis, gall bladder, gall bladder stones
10	Cancer	Immune stimulation, hyperthermia, antitoxin, thymus extract, sono-photo dynamic, bio-electromagnetics, chemotherapy, radiation, immunosuppression, typhlitis, anorexia, malnutrition, neutropenia, sepsis, male pattern baldness, neoplasia, infertility, ovarian failure, neuropathy, cardiac arrythmia, lymphoma, teratoma, leukemia, cardiotoxicity, hepatotoxicity, radiotherapy, fibrosis, tumor, bleeding cancer, lymph nodes, leukemia, carcinoma, breast cancer, colon cancer, prostate cancer, cervical cancer, melanoma, sarcoma, blastoma