ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Prediction of Diseases using Big Data Analysis

M Archana Bakare, Prof. R.V.Argiddi

PG Student, Dept. of Computer Science and Engineering, WIT, Solapur, Maharashtra, India

Associate Professor, Dept. of Computer Science and Engineering, WIT, Solapur, Maharashtra, India

ABSTRACT: Big data can be referred as huge storage of structured, semi-structured and unstructured data, can be used for analysis for extraction of knowledge. Now a day's many social networking users shares their health linked information on web. Such health related information can use for prediction of diseases. Diseases like asthma, high/low blood pressure, diabetes are most prevalent and costly chronic conditions in the world which cannot be cured. However accurate and timely surveillance data can control the diseases. Now current medical dataset can predict emergencies up to certain level but is not able to produce better result. A better another for this is big data. In this paper a new novel method stands introduced which collect results and prediction data from many social networking sites like twitter, Google search. Specialized categories can be made depending on diseases. A generalized method is discussed in this paper which will predict the fore-coming strokes of different diseases depending on data gathered by social networking sites.

KEYWORDS: Big data, Predictive analysis, Access control,Health issues, privacy.

I. INTRODUCTION

Big Data can be referred as a huge repository of structured, semi structured and unstructured data. Now a days social networking sites are becoming a part of life, every third users shares their medical related information on social networking sites, and tries to get remedies for the diseases. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. Infectious disease is a leading threat to public health, economic stability, and other key social structures. Efforts to mitigate these impacts depend on accurate and timely monitoring to measure the risk and progress of disease. Traditional, biologically-focused monitoring techniques are accurate but costly and slow; in response, new techniques based on social internet data such as social media and search queries are emerging. These efforts are promising, but important challenges in the areas of scientific peer review, breadth of diseases and countries, and forecasting hamper their operational usefulness. Diseases like low/High blood pressure, diabetes, asthma are more random and very hazardous. The impacts of such diseases are very dangerous. These diseases are not easily curable.

The impact of such diseases can be reduced by predicting the future stroke timing. It can be possible by analyzing the different types of data collected over those diseases.Gathering disease remains extremely costly in both social and commercial terms. For instance, the majority of global child mortality is due to conditions such as acute respiratory infection, measles, diarrhea, malaria, and HIV/AIDS [1]. Even in developed countries, infectious disease has great impact; for example, each influenza season costs the United States between 3,000 and 49,000 lives [2] and an average of $87 billion in reduced economic output [3]. Effective and timely disease surveillance — that is, detecting, symbolizing, and measuring the incidence of disease — is a critical component of prevention and justification strategies that can save lives, reduce suffering, and minimize impact. Traditionally, such monitoring takes the form of patient interviews and/or laboratory tests followed by a official reporting chain; while generally considered accurate; this process is costly and announces a significant lag between observation and reporting.

There has been increasing interest in gathering non-traditional, digital information to perform disease surveillance. These include diverse datasets such as those stemming from social media, internet search, and environmental data. Twitter is an online social media platform that enables users to post and read 140-character messages called "tweets". It

DOI: 10.15680/IJIRCCE.2016. 0404170

is a popular data source for disease surveillance using social media since it can provide nearly instant access to real-time social opinions. More importantly, tweets are often tagged by geographic location and time stamps potentially providing information for disease surveillance [8,9]. Another notable non-traditional disease surveillance system has been a data-aggregating tool called Google Flu Trends which uses aggregated search data to estimate flu activity [10,11]. Google Trends was quite successful in its estimation of influenza-like illness. It is based on Google's search engine which tracks how often a particular search-term is entered relative to the total search-volume across a particular area. This enables access to the latest data from web search interest trends on a variety of topics, including diseases like asthma. Air pollutants are known triggers for asthma symptoms and exacerbations [12]. The United States Environmental Protection Agency (EPA) provides access to monitored air quality data collected at outdoor sensors across the country which could be used as a data source for asthma prediction. Meanwhile, as health reform progresses, the quantity and variety of health records being made available electronically are increasing dramatically [13]. In contrast to traditional disease surveillance systems, these new data sources have the potential to enable health organizations to respond to chronic conditions, like asthma, in real time. This in turn implies that health organizations can appropriately plan for staffing and equipment availability in a flexible manner. They can also provide early warning signals to the people at risk for asthma adverse events, and enable timely, proactive, and targeted preventive and therapeutic interventions. Our research objective is to leverage social media, internet search, and environmental air quality data to estimate ED visits for asthma in a relatively discrete geographic area (a metropolitan area) within a relatively short time period (days). To this end, we have gathered asthma related ED visits data, social media data from Twitter, internet users' search interests from Google and pollution sensor data from the EPA, all from the same geographic area and time period, to create a model for predicting asthma related ED visits. This work is different from extant studies that typically predict the spread of contagious diseases using social media such as Twitter. Unlike influenza or other viral diseases, asthma is a non-communicable health condition and we demonstrate the utility and value of linking big data from diverse sources in developing predictive models for non-communicable diseases with a specific focus on asthma.

II. RELATED WORK

The Twitter of Babel: Mapping World Languages through Micro blogging Platforms Delia Mocanu, Published: April 18, 2013

In this he shows that data assembled from mobile devices founds to be more accurate than surveys and physical sources. While classical data sources, such as surveys or census, have a limited level of geographical resolution or are restricted to generic workdays or weekends, the data coming from mobile devices can be precisely located both in time and space. Most previous works have used a single data source to study human mobility patterns. Data is collected from three different springs: Twitter, census, and cell phones. The analysis is focused on the urban areas of Barcelona and Madrid, for which data of the three types is available. They have assessed the association among the datasets on different aspects: the spatial scattering of people concentration, the temporal evolution of people density, and the mobility patterns of individuals. Their results show that the three data sources are providing comparable information. Even though the representativeness of Twitter relocated data is lower than that of mobile phone and census data, the connections between the population density profiles and movement patterns detected by the three datasets are close to one in a grid with cells of 2×2 and 1×1 square kilometers.

Social media have been predicted as a data source for flu surveillance because they have the potential to offer real-time access to millions of short, geographically localized messages covering information regarding personal well-being. However, accuracy of social media surveillance systems declines with media attention because media attention increases "chatter" – messages that are about flu but that do not pertain to an actual infection – covering signs of true flu prevalence. This paper summarizes our recently developed influenza infection detection algorithm that automatically categorizes relevant tweets from other chatter, and we describe our current influenza surveillance system which was actively positioned during the full 2012-2013 flu season. Our objective was to analyze the performance of this system during the most recent 2012–2013 flu season and to analyze the performance at multiple levels of geographic granularity, unlike past studies that focused on general or regional surveillance. Our system's influenza prevalence estimates were strongly correlated with surveillance data from the Centers for Disease Control and Anticipation for the United States ($r = 0.93$, $p < 0.001$) as well as investigation data from the Department of Health and Mental Hygiene of New York City ($r = 0.88$, $p < 0.001$). Our system detected the weekly change in direction (increasing or decreasing) of flu prevalence with 85% accuracy, a nearly twofold increase over a simpler model, demonstrating the utility of explicitly distinguishing infection tweets from other chatter.[21]

The US health care system is rapidly adopting electronic health records, which will dramatically increase the quantity of clinical data that are available electronically. Simultaneously, rapid progress has been made in clinical analytics—techniques for analyzing large quantities of data and gleaning new insights from that analysis—which is part of what is known as big data. As a result, there are unprecedented opportunities to use big data to reduce the costs of health care in the United States. We present six use cases—that is, key examples—where some of the clearest opportunities exist to reduce costs through the use of big data: high-cost patients, readmissions, triage, adverse events, and treatment optimization for diseases affecting many organ systems. We discuss the types of insights that are likely to emerge from clinical analytics, the types of data needed to obtain such insights, and the infrastructure—analytics, algorithms, registries, assessment scores, monitoring devices, and so forth—that establishments will need to perform the necessary analyses and to implement changes that will improve care while reducing costs. Our encounters have policy implications for regulatory oversight, ways to address privacy sufferings, and the support of research on analytics

Culotta, Aron. "Towards detecting influenza epidemics by analyzing Twitter messages." In Proceedings of the firstworkshop on social media analytics, pp. 115-122. ACM, 2010.[14]

In This proposed method based on analysis of messages posted on the micro-blogging site Twitter.com to determine if a similar connection can be uncovered. They proposed several methods to identify influenza-related messages and equate a number of regression models to correlate these posts with CDC statistics. Using over 500,000 messages spanning 10 weeks, we find that our best model achieves a correlation of .78 with CDC statistics by leveraging a document classifier to identify relevant messages.Analyzing user messages in social media can measure different population features, including public health measures. For example, recent work has correlated Twitter messages with influenza rates in the United States; but this has largely been the extent of mining Twitter for public health. Michael J. Paul considered a broader range of public health applications for Twitter. They apply the newly introduced Ailment Topic Aspect Model to over one and a half million health related tweets and discover mentions of over a dozen ailments, including allergies, obesity and insomnia. They introduced postponements to include prior knowledge into this model and apply it to several tasks: tracking illnesses over times (syndromic surveillance), measuring behavioral risk factors, localizing illnesses by environmental region, and analyzing symptoms and medication usage. Their results suggest that Twitter has broad applicability for public health research.

Centers for Disease Control and Prevention, —About the Morbidity and Mortality Weekly Report (MMWR) Series ‖ [Online] Available: http://www.cdc.gov/mmwr/about.html

Syndrome surveillance, the monitoring of clinical syndromes that have significant impact on public health, impacts medical resource allocation, health policy and education. Many common diseases are continuously monitored by collecting data from health care facilities, a process known as sentinel surveillance. Resources limit surveillance, most especially for real time feedback. For this reason, the Web has become a source of syndromic surveillance, operating on a wider scale for a fraction of the cost. Google Flu Trends (Ginsberg et al. 2008) tracks the rate of influenza using query logs on a daily basis, up to 7 to 10 days faster than the Center for Disease Control and Prevention’s (CDC) FluView (Carneiro and Mylonakis 2009). High correlations exist between Google queries and other diseases (Pelat et al. 2009), including “Lyme disease” (Seifter et al. 2010). These results fall under the area of infodemiology (Eysenbach 2009). Similar results exist for Twitter, which can be a complimentary resource to query logs, but may also contain freely available and more detailed information; people write detailed messages for others to read. Lampos and Cristianini (2010) and Culotta (2010b) correlated tweets mentioning the flu and related symptoms with historical data. Similarly, Quincey and Kostkova (2010) collected tweets during the H1N1 pandemic for analysis. Ritterman, Osborne, and Klein (2009) combined prediction markets and Twitter to predict H1N1. More generally, Chew and Eysenbach (2010) evaluated Twitter as a means to monitor public perception of the 2009 H1N1 pandemic. Scanfeld, Scanfeld, and Larson (2010) evaluated the public understanding of antibiotics by manually reviewing Tweets that showed incorrect antibiotic use, e.g., using antibiotics for the flu. The public health community is also considering how social media can be used to spread health information, with applications including risk communication and emergency response[24].

DOI: 10.15680/IJIRCCE.2016. 0404170

Vance, Howe, and Dellavalle (2009)

analyzed the pros and cons of using social media to spread public health information in young adults. Pros include low cost and rapid transmission, while cons included blind authorship, lack of source citation, and presentation of opinion as fact. Greene et al. (2010) studied how medical information is exchanged on Facebook, where groups specific to diseases share information, support, and engage patients in their diseases[25]

Review of Extracting Information from the social web for health personalization

viewed www.ncbi.nlm.nih.gov › NCBI Literature PubMed Central (PMC) Published online 2011 Jan28.Luis Fernandez-LuqueNorthern ResearchInstitute, Postboks 6434 This paper provides a review ofdifferent approaches for extracting information from the Social Web for health ......Bonander J. Personally tailored health information: a health 2.0 approach.also different approaches for extracting information from social web applications to personalize health care information. The model we use in this paper could be used to analyze tweets for health care personalization. Finally, the public is considering the larger impact of how social media can impact health care, where patients can “friend” doctors and always share information among thousands of friends[22].

III. Proposed System

The aim of this work is to provide analysis of data results and also used for diabetes, low & high blood pressure. Here data is collected from social media, internet search. It is a popular data source for disease surveillance using social media since it can provide nearly instant access to real-time social opinions. More importantly, time stamps possibly providing information for disease investigation. The aim of this work is to provide analysis of data results and also used for diabetes, low & high blood pressure. Here data is collected from social media, internet search. It is a popular data source for disease surveillance using social media since it can provide nearly instant access to real-time social opinions. More importantly, time stamps possibly providing information for disease investigation.

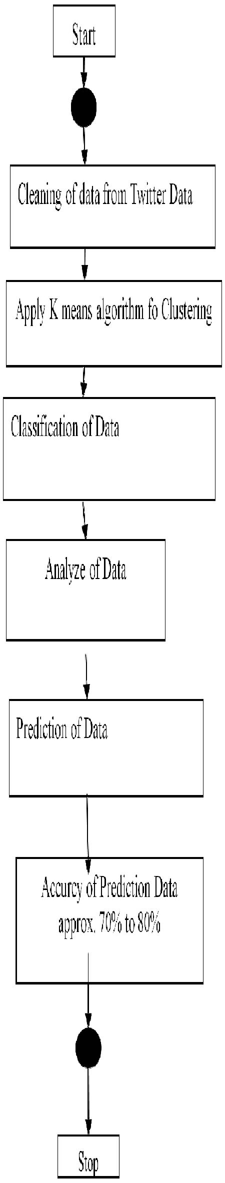
IV. Design and Architecture

Diseases are asthma, high/low blood pressure, diabetes are most prevalent and costly chronic conditions in the United States which cannot be cured. However accurate and timely investigation data can control the diseases and does not allow crossing certain threshold. Now current medical dataset can predict dangers up to certain level but is not able to produce better result. In this paper a new novel method is introduced which collect outcomes and prediction data from many social interacting sites like twitter, Google search and environmental sensor models. Particular categories can be made depending on diseases. Diabetes is widely recognized as one of the leading causes of death and disability. However, diabetes is likely to be under reported as the innovative cause of death on death certificates. About 65 percent of deaths between those with diabetes are official to heart disease and hit.

For diseases like blood pressure, regular control is necessary. This means making lifestyle changes, taking prescribed medicines, and getting continuing medical care. Treatment can help control blood pressure, but it will not cure HBP. If treatment is stopped, blood pressure and risk for related health difficulties will rise.

The results obtained with other diseases like high/low blood pressure and diabetes. The results can be helpful for public health investigation, emergency department preparedness, and, targeted patient interventions.

In that our model can predict the number of diseases stroke visits based on near-real-time environmental and social media data with approximately 70% precision. The results can be helpful for public health surveillance, emergency department preparedness, and, targeted patient interventions. A generalized method is proposed in this paper which will predict the fore-coming strokes of different diseases depending on data gathered by social networking sites. The impact of such diseases can be minimized by predicting the coming stroke timing. It can be possible by analyzing the different types of data composed over those diseases.


Start
Cleaning of data from Twitter Data
Apply K means algorithm fo Clustering
Classification of Data
Analyze of Data
Prediction of Data
Accurcy of Prediction Data approx. 70% to 80%
Stop

Phases of project
1. Collection of data from various social networking sites.
2. Data is cleaned. Unwanted data is removed and transformed usefulness of application.
3. Clusters can be created using k means algorithm. Depending on the number of diseases found in communication.
4. For clustering k means algorithm is used.
5. Multirank walk algorithm is used for classification of data.

Multirank walk algorithm is used for prediction of disease fore coming depending on data collected from social networking site. This prediction can used to take preventative measures for patients and can be used for treatment purposeBig data is source of structured, unstructured and semi-structured data. Efficiency of clustering algorithm depends on relevancy of data found. Prediction of decision depends on clustering, data extraction and data cleaning. A good cluster can result better prediction result.

Data Collection Phase:

The data underlying our studies has been collected using the social web search engine Topsy (http://www.topsy.com) and Twitter API (http://apiwiki.twitter.com). For selecting tweets, a list of symptoms and disease names has been created manually by domain experts. This list contains symptoms of infectious diseases such as fever, headache and disease names such as asthma, swine flu or H1N1 (their German correspondents). For the relevance assessment of Twitter messages , only this list has been used for filtering and selecting the Twitter messages for the evaluation.

Data Analysis:

(Relevance Assessment of Tweets), tweets matching a disease name or symptom are manually classified as relevant or irrelevant for disease investigation. Any tweet should be labeled as positive or case regardless whether it is a established, supposed or possible case, if:

1. It confirms that the user is infected with a disease or symptom, e.g, I am sick now... I got asthma and I need medicine,

2. It confirms that another subject (e.g., animal,) has a disease or symptom,

3. A test result is stated which confirms an infection, e.g.

Tyler is asthma positive! or if

4. A doubt is mentioned, e.g., my son is assumed to has asthma, or

5. Another outbreak or danger is described.

V. Conclusion

The Data mining techniques can be used to extract the useful information from big data. A Multirank algorithm can be used for prediction of asthma. Accuracy of process depends on relevant found data sssin big data. 80 % accuracy can be achieved in prediction of diseases.

References

1. Predicting Asthma-Related Emergency Department Visits Using Big Data Sudha Ram, Member, IEEE, Wenli Zhang, Max Williams, and Yolande Pengetnze, MD
2. Peymanfar, A. Khoei and Kh. Hadidi, "A New ANFIS Based Learning Algorithm for CMOS Neuro-Fuzzy Controllers Electronics", 14th IEEE International Conference on Circuits and Systems, ICECS, 2007, pp.890-893.
3. K. P. Adlassnig, "Fuzzy set theory in medical diagnosis," IEEE Trans. Syst., Man Cybern., vol. 16, no. 2, pp. 260-265, Mar. 1986.
4. C. D. Stylios and P. P. Groumpos, "Modeling complex systems using fuzzy cognitiv maps," IEEE Trans. Syst., Man Cybern., Part A: Syst. Humans, vol. 34, no. 1,pp. 155-162, Jan. 2004.
5. R. I. John and P.R. Innocent, "Modeling uncertainty in clinical diagnosis using fuzzy logic," IEEE Trans. Syst., Man, Cybern., vol. 35, no. 6, pp. 1340-1350, Dec. 2005.
6. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ (2006) Global and regional burden of disease and risk factors, 2001: Systematic analysis of population health data. The Lancet 367: 1747–1757.
7. Thompson M, Shay D, Zhou H, Bridges C, Cheng P, et al. (2010) Estimates of deaths associated with seasonal influenza — United States, 1976–2007. Morbidity and Mortality Weekly Report 59.
8. Molinari NAM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, et al. (2007) The annual impact of seasonal influenza in the US: Measuring disease burden and costs. Vaccine 25: 5086–5096.
9. Broniatowski, David A., Michael J. Paul, and Mark Dredze. "National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic." PloS one vol. 8, no.12, e83672, 2013.
10. Kim, Eui-Ki, et al. "Use of hangeul twitter to track and predict human influenza infection." PloS one vol. 8, no.7, e69305, 2013.
11. Paul, Michael J., and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." In ICWSM, pp. 265-272. 2011
12. Fox, Christopher. "A stop list for general text." In ACM SIGIR Forum, vol. 24, no. 1-2, pp. 19-21. ACM, 1989.
13. Krieck, Manuela, Johannes Dreesman, Lubomir Otrusina, and Kerstin Denecke. "A new age of public health: Identifying disease outbreaks by analyzing tweets." In Proceedings of Health Web-Science Workshop, ACM Web Science Conference. 2011.
14. Culotta, Aron. "Towards detecting influenza epidemics by analyzing Twitter messages." In Proceedings of the first workshop on social media analytics, pp. 115-122. ACM, 2010.
15. Peymanfar, A. Khoei and Kh. Hadidi, "A New ANFIS Based Learning Algorithm for CMOS Neuro-Fuzzy Controllers Electronics", 14th IEEE International Conference on Circuits and Systems, ICECS, 2007, pp.890-893.

16. Akinbami LJ, Moorman JE, Liu X. “Asthma prevalence, health care use, and mortality: United States, 2005—2009”. National health statistics reports no. 32. Hyattsville, MD: National Center for Health Statistics; 2011.

17. Centers for Disease Control and Prevention. "Vital signs: asthmaprevalence,diseasecharacteristics, and self-management education: United States, 2001--2009." MMWR. Morbidity and mortality weekly report vol. 60, no.17 pp. 547, 2011.

18. National Institutes of Health. "Guidelines for the Diagnosis and Management of Asthma." Expert panel report. vol. 2. NIH, 1997.

19. Centers for Disease Control and Prevention, “About the Morbidity and Mortality Weekly Report (MMWR) Series” [Online] Available: http://www.cdc.gov/mmwr/about.html

20. Pesola, Gene R., Feng Xu, Habibul Ahsan, Pamela Sternfels, Ilan H.Meyer, and Jean G. Ford. " Predicting asthma morbidity in Harlem emergency department patients." Academic emergency medicine vol.11, no.9, pp. 944-950.

21. The Twitter of Babel: Mapping World Languages throughMicroblogging PlatformsDelia Mocanu, Published: April 18, 2013

22. Review of Extracting Information from the social web for health personalization viewed www.ncbi.nlm.nih.gov › NCBI Literature PubMed Central (PMC) Published online 2011 Jan 23. Analyzing Twitter for Public Health Michael J. Paul and Mark Dredze Human Language Technology Center of Excellence Center for Language and Speech Processing Johns Hopkins University Baltimore, MD 21218 {mpaul,mdredze}@cs.jhu.edu

23. Centers for Disease Control and Prevention, ―About theMorbidity and Mortality Weekly Report (MMWR) Series‖[Online] Available: http://www.cdc.gov/mmwr/about.html

24. Social internet sites as a source of public health information. www.ncbi.nlm.nih.gov/pubmed/192546562009 Apr;27(2):133-6, vi. doi: 10.1016/j.det.2008.11.010. Social internet sites as a source of public health information. Vance K(1), Howe W, Dellavalle RP.

BIOGRAPHY

Ms. Archana Bakare is a pursing Master of Engineering in the Computer Science Department, in Walchand Institute of Technology, Solapur.She received Bachelor of computer Engineering degree in 2013 from Solapur Univercity, Solapur.

Prof. R.V.Argiddi is Associate Professor in Computer Science in Walchand Institute of Technology, Solapur.He is H.O.D in Walchand Institute of Technology, Solapur From Computer Science Department.