ANALYSIS OF PUBLIC HEALTH INFORMATION FROM SOCIAL MEDIA

By

Madhav Narayan Thalambedu Kumaresan

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
In partial fulfillment of the requirements
For the degree of Master of Science in
Information Technology

Charlotte

2016

Approved by:

____
Dr. Mohsen M Dorodchi

____
Dr. Lawrence E Mays

____
Dr. Daniel A Janies

____
Dr. Heather R Lipford

ABSTRACT

MADHAV NARAYAN THALAMBEDU KUMARESAN Analysis of public health information from social media. (Under the direction of DR MOHSEN M. DORODCHI).

Social networking sites have become an integral part of daily life in the modern world. They can serve as an important, dynamic data source providing collective intelligence and awareness of health related issues. Analysis of huge volumes of unstructured data from social media for useful information could be a challenging job due to different factors. This work focuses on collecting data from various publicly available sources, applying data cleansing methods on collected data and analyze the extracted data. This study deals with the creation of methods to scrape data from different social media websites, followed by preparing and cleaning of data involving tasks like stemming of words. Finally, this study applies ontology analysis to find the co-occurrences of keywords of interests to measure associative strength between them. It further computes the corresponding support and confidence intervals to form rules.

The summary of the procedures to extract data and preparing them for analysis is produced as results of this work. Our findings suggest that such analysis requires engagement of domain expert from the launch point of the research. This is due to the fact that enormous number of experiments needs to be executed and analyzed to be able to extract useful information . Therefore we can conclude that social media websites such as Facebook are very critical sources of data for analysis of social aspects of health, especially outbreak. We also observe a wide range of discussion about the diseases that occur, across different regions of the globe. These observations can help public health officials to identify and follow social impact on disease outbreaks, so that

publicly appropriate actions can be recommended. We also find that social media can be used as a source of information for health related research as well as to discover patterns that are more interesting to researchers across higher education institutions and universities.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCING THE PROBLEM

Social media has emerged as a platform for multi-dimensional interactions, with the ability to share information, shape opinion, and connect individuals and communities along with the capacity to increase participation (2014, Maggiani). Proponents of social media have been increasing on a yearly basis. Pew Internet[1] reported there has been a tenfold increase in the use of social media in the past decade. Nearly two-thirds of American adults, which amounts to 65 percent of the population, use social networking sites, increasing from only 7 percent in 2005. Osterrieder (2013), in her research, mentioned that social networking sites like Facebook and Twitter allow users to share and interact with like-minded people. She also mentioned its strength as rapid dissemination and amplification of content with the ability to lead informal conversations, making it a powerful tool for use in a professional context for data analytics.

The Internet as a whole is an important resource for health related information exchange. However, more recently, it is observed that people communicate their health-related concerns more frequently on social media. In a research on influenza surveillance, Polgreen (2008) mentioned that frequency of internet searches could yield information regarding infectious disease activity. It was observed that there exists a “distinct temporal

[1] www.journalism.org/2015/07/14/news-habits-on-facebook-and-twitter/

association" between influenza disease activity and influenza related search term frequency (Polgreen, et al. 2008). As it is reported, only in the US, the search for the term "Influenza" had a positive correlation (with p value < 0.001) with the number of deaths due to pneumonia and influenza within five weeks of the first reported case. These and similar studies reveal that there exists an underlying pattern between disease outbreaks in particular (and health information in general) and the information communicated across social media. Rakesh, et al. (1997) in his classical method for identifying relationships between variables using the measure of interestingness (Matheus, et al. 1993). They introduced the concept of association rules for discovering regularities between objects in a large-scale transaction database in super markets. Even though the focus of this work is mainly on data retrieval, this method has the potential to be extended to identify the associations between disease outbreaks and their locations from social media.

The proposed methodology examines the challenges faced during data gathering, data cleaning, and data analysis phases. Moreover, it uses association strength between diseases and the corresponding locations to predict support and confidence values. To analyze data from social media, a corpus focusing on varied communities in Facebook was created with the emphasis on news channels. Cleaton (2015) proved that information derived from the systematic collection i.e. WHO[2] reports and analysis of unstructured news reports provided by authoritative sources (news channels) is a reliable means to assess epidemiological patterns of evolving infectious disease threats with high confidence. This guided us to choose news channels like CNN, ABC, and BBC, as well as public news

[2] www.who.int/en/

forums like Buzzfeed[3] and Humans of New York[4].as our raw data sources. In order to focus on public health news, organizations such as NIH, CDC, and Infectious Disease News[5] were included as well.

Data collected from these communities were organized into a single documented corpus, which is further stripped for white spaces, punctuations, followed by stemming of words to remove suffixes of verbs and unify words from the same roots. In addition, procedures for removing words that are identified as noise were carried out, resulting in data that had plain text without any symbols or punctuations and ready to be processed. This data was split and analyzed for frequent terms to arrive at the most frequently used words in the corpus.

In a research by Kibbe, et al., (2015), focused on an expanded and updated database of human diseases for linking biomedical knowledge through disease data. They proposed a method of integrating disease terms from various systems, including MeSH[6], ICD-9[7]and NCI[8] to form a repository with ontologies of diseases across the globe. We used their repository, and compared the disease ontologies with the frequent words from the corpus and arrived at a list of words (diseases) and their frequencies. This result gives a picture of the actual diseases that are in place during that particular period. On further analysis of a particular disease, we can compute the support and confidence measures that arises due to the co-occurrence of diseases and locations. In future, the method can be extended derive information regarding the origin and spread of diseases in different countries.

[3] www.buzzfeed.com/

[4] www.humansofnewyork.com/

[5] www.outbreaknewstoday.com/

[6] meshb.nlm.nih.gov/#/fieldSearch

[7] icd9cm.chrisendres.com/

[8] ctep.cancer.gov/protocolDevelopment/codes_values.htm

This thesis report is organized as follows. The following section surveys the current and past methods, focusing on aspects like data collection, cleaning, and analysis. The corresponding section explain the different tasks that were carried out in identifying the literature and the methodology implemented in the process.

1.1. Literature Survey

The initial part of the research focused on finding literature relevant to different forms of data gathering. Technological advancements have provided new forms of socialization and information exchange, with web 2.0 and virtual worlds helping to streamline education systems (Harris and Rea, 2009). There are multiple channels of information that are generated through communities with the goal of better distribution and helping to connect like-minded people (Parisa, 2010). Social media has also replaced many traditional methods of communication, such as newspaper and television. With the increase in the number of socially active participants, there are huge sets of data generated with an enormous potential to analyze data and mine patterns. According to Pew Research Centre and American Trends panel[9], nearly 54 percent of the U.S. users of Facebook see some news regarding health and medicine and about 28 percent of Facebook users post information about the news.

Marjolijn (2013) showed that 52.3 percent of patients use Facebook as a means of seeking social support and exchanging advice. Since the participants are not completely anonymous, it is believed that there exists a feeling of increased engagement, accountability, and more responsibility than stand-alone communities. In establishing an

[9] www.journalism.org/2015/07/14/news-habits-on-facebook-and-twitter/

automated source for surveillance, Keller and M., Freifeld (2009) created an automated vocabulary for Geo-Parsing online epidemic intelligence and helped in improving the existing work on HealthMap[10] (Barton and Grant, 2006). This was accomplished by training data sets generated with the help of statistical machine learning approaches and was able to reduce the number of false positives. This helped us to understand how information from social media can be mined for extracting useful information.

However, there are few academic reports of research that study the efficiency of using social media in public health analytics and prediction. Most of these point toward the analysis of patient health records based on repositories collected by public health organizations. There were studies that aimed at successfully predicting diseases from trivial repositories of news channels that are static, but there were very few studies that are based on the dynamic interactions presented on social media platforms. To elaborate a news report once aired cannot be altered compared to social media where posts can be altered at any point in time.

In this chapter, we review the current status of academic literature. First, an analysis of patient health records is reviewed with the emphasis on extraction of medical information from repositories. This is one of the key features behind pattern mining. Xiaohua zhou and Hyoil Han (2006) proposed a Medical Information Extraction (MedIE) system to analyze patient information from free text medical records. They combined an ontology based approach for information (medical terms) extraction, a graph based approach using parsing results of link-grammar parser and a NLP based feature extraction method with ID3[11] based decision tree to perform text classification. This methodology

[10] www.healthmap.org/en/

[11] www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm

was able to predict and identify diseases with the highest level of accuracy when compared with consultation notes.

One of the first steps when handling textual data is the extraction of relevant information from a large collection. We concentrate on extracting disease information from a large repository of unstructured data. Ralp and Roman (2002) present an analysis of a system called Proteus-BIO, which provides the capability to search documents about infectious diseases from the Web. The framework accumulates Web pages, extracts data about diseases, and exhibits the extracted data in a tabular structure with connections back to the webpages. The major components of any information extraction system is a web crawler, which scrapes information from web pages, an extraction engine that consolidates data and forms a data repository and finally a framework, which allows users to search the repository and gather relevant information. Ananiadou, et al., (2006) describe various steps involved in text mining for knowledge discovery in the context of biomedical literature. Information retrieval, information extraction and data mining play an important part of any data analysis and, recognizing biological entities in text is difficult due to the inconsistencies in naming biological entities.

Kim et al. (2012), emphasize the need for the creation of a persistent and shareable repository to tackle the issues that arise due to active developments in corpora and annotations. This helped us in understanding the use of annotations in helping researches to find topics with ease. Terms are the backbone for biological entities as they reveal specialized knowledge about a particular disease outbreak. Diseases and symptoms often denote themselves in a variety of ways to bring out the same concept, similar to biological term specification, for example, the metabolite glucose-6-phosphate is referred to as variants and permutations of α or β, d- or l-glucose (or hexose)-6-(mono)-phosphate. Furthermore, a term can be given in an extended compounded form then expressed through various mechanisms, including orthographic variations usage of slashes and hyphens e.g., Löfgren-syndrome and Lofgren syndrome), lower and upper cases (H1N1 and H1n1), spelling variations (tumor and tumour), various Latin and/or Greek transliterations (oestrogen and estrogen), and abbreviations (RAR and retinoic acid receptor). Further complexity is introduced when authors vary the forms they use in different ways (e.g. different reductions, such as thyroid hormone receptor and thyroid receptor). Therefore, a term can be viewed as a class of different terminologies.

Hripcsak and Friedman (1995) created a system to show physicians and natural language processors are close in identifying diseases from narrative reports, and natural language processing has the potential to extract clinical information from narrative reports to support automated decision making and clinical research. Yli-Hietanen and Niiranen (2009), in their work prove that domain specific language modeling when implemented as a case specific rewriting system is a highly promising tool for computational understanding of a specific class of medical texts with a focus on records obtained from patients. Protecting the user's identity is a major concern in any analysis. By implementing false positive filtering, it was observed that de-identifying names of patients had a high recall rate and it helped eliminate false positives achieving an F-Score[12] of 92.6 percent (Fernandez and Shen, 2012). Focusing on methods to increase the prediction is always important in analysis.

[12] The harmonic mean of precision and recall where precision is the ratio of true positives to all positives, while recall is the ratio of true positives to all that were classified correctly

Predicting on a corpus for diseases through machine learning models would yield better results compared to dictionary based and rule based models (Fu and Batista, 2015). In their research on automated data acquisition for heart failure (ADAHF) project, they aimed at automatically extracting heart failure treatment performance metrics from clinical narrative documents. It was verified that combining a rule based and a machine learning based system yielded prediction rates with a 99.3 percent recall and 98.8 percent precision. Thus, combining methods across should yield high rates of accuracy (Myestre and Garvin, 2015).

Creating a corpus for analysis of medical data is the starting point for this research. Setting up a clinical data warehouse for analysis involves transforming data from a comprehensive computer-based patient record system (CPRS) into a data warehouse server. A dataset for analysis is created by extracting and cleaning selected variables, and mining of the data using exploratory factor analysis (Parther and Hammond, 1997). Consumer Health Vocabularies are very important in finding expressions and thoughts about health topics. Often it is observed consumers face difficulties during exploration of health vocabularies due to the robustness, societal, and cultural association (West and Hammond, 2015). Hence a vocabulary by a concerted, interdisciplinary and open access approach, by the open access community to help in understanding the overall phenomenon must be formed. Obtaining these records requires several levels of clearances as the data deals with clinical patient records.

News and media are one of the key sources of information and reliable sources. However, based on a research by Moynihan et al., (2000), news media can include inadequate or incomplete information. When analyzing the benefits, risks, and costs of certain drugs they found that there were discrepancies due to financial ties between study groups or experts and pharmaceutical manufacturers. Therefore, researchers have been looking into resources with more caution for extracting reliable information. Chan et al., (2010) analyzed the entire public news records of WHO to find the changes in the communication process of an outbreak. They found from year to year the media was discussing disease. In a research on the use of n-grams and semantic features for the classification of outbreaks from online news, Conway et al., (2009) proved that there is an increase in the accuracy of information when running a classification algorithm with a tagger. This clearly helps us to understand the reliability behind the use of news as a medium of health related information. Riga et al, (2014) analyzed the relationship between tweets linked to media reports on public health. They used the relationships between social media content and real-time observation for urban air quality and public health. This analysis proves that integrating social media with news reports and public health yields better options for analysts to increase the predictions. These features of news as a source of information ensured that news related social media posts was a part of future analysis.

To arrive at a corpus of data with an unrestricted flow of information from a global community, we focus on social media. Social media as a whole is a platform for users from all over the globe to share their reactions with minimum restriction. The enormous amount of information accompanied with high frequency of changes, encourages forming of extensive customer groups, which serves as an attractive open entry for harnessing data into a structure that considers specific estimates about particular results, without the necessity for a complex framework. One can also build models to aggregate opinions of a collective group and gain useful insights into their behavior, while predicting future trends

(Asur and Huberman, 2010). Gathering information on how people converse regarding particular products can be helpful when designing marketing and advertising campaigns. Designing a healthcare data model using the traditional top down model of data warehousing is not fruitful[13]. The goal of creating a data corpus varies, as the focus is not dependent on a particular motive, in contrast with conventional methods. We determine everything in advance to be able to analyze to improve outcomes, safety and patient satisfaction. Creating a data store by integrating data from different sources is a daunting case. In case of setting up an individual data mart, which is an ideal scenario for bottom up data model construction, issues like bombarding source system quite frequently, missing granular data for fine-grained analysis and binding data in an early state. The ideal scenario for setting up a data repository would be to implement just in time binding where data is taken in its atomic form with minimal processing and creating a data mart by drawing it from these sources. Finally binding data when necessary, i.e. when a specific use-case or business calls for it this method gives maximum flexibility for using data to tackle wide variety of use cases and reduces wasting of resources. This method would help us to see what is coming down the road in future in contrast to making decisions about the data model up front.

The DARPA (Defense Advanced Research Project Agency) network challenge funded by The United States Department of Defense in 2009, tested the ability of online social networks to mobilize teams and solve a real world problem, which could potentially improve disaster response and coordination of relief efforts. Lerman, K., and Ghosh, R. (2010) advocate that understanding the characteristics of user activity and the underlying

[13] www.healthcatalyst.com/whitepaper/3-approaches-healthcare-data-warehousing

effect networks have on it is essential for the effective use of social media and peer production systems. Since people create social links with others similar to them, the dynamics of information spread might be different from its spread outside the network compared to its spread inside the network. Isolating in-system from out-of-system action permits us to better gauge the intrinsic nature of the commitments (Crane and Sornette 2008) or foresee their future action (Lerman and Galstyan 2008; Hogg and Lerman 2010). Yepes and Han (2015) investigate on analysis of twitter for public health surveillance. Their technique creates a database by extracting and segregating tweets based on geographical locations, which is scrutinized to predict if there are any disease outbreaks. This architecture is analyzed for named entities by a recognizer called Micromed[14]. This was an inspiration for the creation of a corpus with huge unstructured data from social media. Lampos, V. and Cristianini, N. (2010, June) proposed a method of tracking flu epidemic in UK by using the contents of Twitter. This information is an early warning system for government agencies, and serves as an information provider to most of the health institutes to equip themselves with necessary remediation measures. This approach yielded a correlation coefficient of 95 percent when compared with data from public health agencies. Sokolova et al,(2012) in their analysis describe the process of mining personal health information form social media using terms and WordNet, and illustrate the limitations of keyword search and understand the use of ontologies to get a better outlook of what people are discussing . Paul et al,(2011) in their research on analyzing twitter for public health. They extend the ailment topic aspect model incorporating concepts like features like syndromic surveillance, measuring behavioral risk factors, localizing illnesse

[14] www.micromed.eu/

and analyzing symptoms and medication usage. They compute quantitative and qualitative correlations between public health data and twitter feeds to emphasize the use and limitations of twitter based analysis. Eichstaedt et al,(2015) studied twitters capacity to predict county level heart disease morality based on psychological language used in tweets. The group found that topic based models for analyzing tweets is more efficient than the dictionary based approach . This study was one of the key factors implemented during the stemming of words where words are grouped based on topics after classification. In a study (Oh, et al., 2013), conducted among 291 users who used Facebook as a medium for seeking Health Related Social Support. There was a positive association between health concerns, seeking, perceiving health related social support and enhanced self-efficacy. This proves that Facebook can become a platform for discussion among individuals all over the globe on health related issues. Prieto et al, (2015) proposed an automated method of data extraction and analysis from tweets, taking the benefits of the wealth of data provided by twitter to measure the incidence of a set of diseases in society. The method consists of two steps: The first step focuses on filtering tweets with specifically crafted regular expressions, and the second step, consists of manually labeling tweets as positive or negative that is used to train classifiers. Once accurate filters and classifiers are created, we can proceed with analyzing the incoming stream of data for measuring health incidences. This method is generic and could be easily adapted to other languages or geographical areas and to other health and social conditions by creating regular expressions following the same procedure and by training new models for classifications.

Lamprecht, et al, 2014, in their research on using ontologies to model human navigation, to address navigation in social networks through a decentralized algorithm to find a way to solve the pathfinding problem in social networks. The term decentralized stems means, the search proceeds by forwarding the problem from one node to another. Decentralization is applied in a social network represents a person taking decision at each node. One of the most prominent concepts involved in information networks is "information foraging" (Priolli et al, 2007). In this theory, there is no relationship between background knowledge and information networks; instead, it is catered by information scent, which is unique and dependent on the search target. For instance, when searching for information on camels, a link leading to an article about Sahara Desert would provide more scent (information), rather an article about penguins at Antarctica. This concept is very useful when we try to map ontologies on a huge repository of data. This would help us in mapping the association between different terms with each term representing a node.

CHAPTER 2: RELATED LITREATURE WORK

Data is retrieved from web pages by the mechanisms of browsing and keyword searching, these mechanisms are inefficient for locating particular items of data because following links is tedious and it is easy to get lost. Keyword searching can be efficient but often they return vast amounts of data that cannot be handled. In spite of data being public and readily available for consumption, it is often difficult to query the web and manipulate information as done in the case of databases. Database require structured data so a traditional database technique cannot be applied. The advent of XML as a global standard for web application helped to decrease the problem, but the volume of unstructured or semi-structured data available on the web is huge. Thus, to tackle this problem a possible strategy is to extract data from web sources and populate databases (Alberto et al, 2002). The process of extracting data from websites involves writing specialized programs called wrappers that classifies the data of interest and maps them to a suitable format. The most challenging feature of a wrapper is to identify the content of interest among the other uninteresting pieces of text, and the structural variations exhibited across various sources. The solution for these problems must be to develop a tool that is highly accurate and robust, while demanding a little effort from the developers. Therefore, we used the RFacebook package to span across different Facebook pages based on their ids and extract the content of the posts from users. Prerequisites for this task included creating and registering an Open Authentication key, which grants permission to scrape data from public pages. Data extracted is deposited in a file with a comma separated value extension, and a corpus of data is created.

Cleaning of data is one of the important tasks before analyzing data, which requires a series of steps to be performed during the process. This task starts with the creation of the unified repository of data. Data in the corpus is accumulated across different repositories, and must be purged of discrepancies and transformed into a uniform format before it can be used for analysis. The Data cleaning process should satisfy several requirements. The first task of a good cleaning mechanism should be to detect and remove all major errors and inconsistencies in an individual data source as well as an integrated data source. Various functions involved in data cleaning on one source must be reusable for other data sources. In case of large data sources, a workflow that supports and executes all data transforms from multiple sources as well as on a large data set must be designed in an efficient way. Taking all these factors into consideration we have created modules for stripping white spaces, the spacing between words are sometimes added to make data more presentable, since we are concerned only about the raw data, these spaces might sometimes act as a noise in the data unnecessarily adding complexity for analysis. This is followed with non- text filtering, where the module is coded to detect tags and signatures in the post, also this module is responsible for removing special characters (example: *, >, <=, many other similar symbols) that add complexity to the text. It also removes symbols (including space, exclamation mark, question mark, and period) at the sentence or word ending. Furthermore, it eliminates tokens (example: non-ASCII characters, tokens containing many special symbols, and lengthy tokens). This module is followed by removing stem words which are words ending with "-ing" and "-es" along with a list of words available as a part of an inbuilt dictionary. Followed by this we remove words that do not necessarily help in analysis, this manual task involves grouping of words, which does not have a meaning. For an instance, when we are analyzing data from Twitter, the word Twitter becomes insignificant because, it is generic and mostly available across all the posts. Similarly, the word google becomes insignificant when we are analyzing data from google pages. These insignificant words are termed as stop words, their major task is to add grammatical meaning to the sentences and they act as a noise when we are making a real time analysis. Stop words are generally thought to be a single word phrases but contextually it differs based on the applications. For example, removing some words like “the”, “a” and “an” known as determiners, prepositions like “above”,” across” and “before” and removing some adjectives like “good” and “nice” can be appropriate. For some applications that deal with complex sentimental analytics this can prove detrimental. For instance, while performing sentimental analytics removing adjectives like ‘good’ and ‘nice’ as well as negations such as “not” can lead an algorithm off track. In these cases, choosing minimal stop lists consisting of just determiners or determiners combined with prepositions or just coordinating conjunctions depending on the needs of the applications are very important. Statements that appear frequently might pose a threat to the analytics, for instance, if we find a phrase like “good item” appearing frequently in our corpus but has a very low discriminating power this might lead to an unwanted behavior in our results. One way to avoid this is to exclude the phrases as a whole which can be coined as stop phrases. It is possible to construct such phrases that occur very frequently in our document and reduce the variations caused due to these phrases.

Creating a domain specific stop words list is very important to tailor the list of phrases or words that must be excluded in order to avoid variations in analysis. One simple way to do it is to measure the sum of frequencies of unique word that we come across the corpus, sort them in descending order of term frequencies, and take the top N terms as stop words. We can also eliminate common English words prior to sorting to target domain specific words. The benefit of this approach is, it is very simple and easy to implement and in case if there is a huge document which is dominant and causes the terms to be at top we can normalize the document with term frequencies using the document length. We can split the document based on the size, create a separate document, and use it for analysis. Another aspect to be considered while constructing a domain specific stop words list is the least specific words in context, for instance, words like username “@username” which is very common in a collection of tweets or posts is not useful in general. Some other terms or phrases like “TTYL” which is a user made acronym is also not beneficial during a context specific analysis, to avoid the problem we must also consider these infrequent terms in our analysis. The final method is to compute the inverse document frequency. The two approaches above suffer a critical problem; all terms are considered equally important when it comes to frequencies (Christopher et al, 2008). For example, considering the example of a collection of documents relating to an auto industry there are repeated occurrences of words like insurance. An idea would be to scale down the terms with high frequencies. This would give us an idea of the occurrence of terms across the corpus irrespective of the number of documents. This measure is often un-even when considering different measures of term distributions across each document. To scale it more reasonably we use a term called inverse document frequency where we reduce the weight of a term by a factor that grows with its collection frequency. Intuitively we are providing a higher boost to the words that have low frequencies and try to reduce the values for words having very high frequency. This helps in including words and terms with low frequencies in the analysis.

The final task to be implemented in this process is conversion of letters into lower case, which reduces the discrepancies due to character case variations. As a last step, we ensured numbers are removed from the posts in order to reduce the discrepancies due to

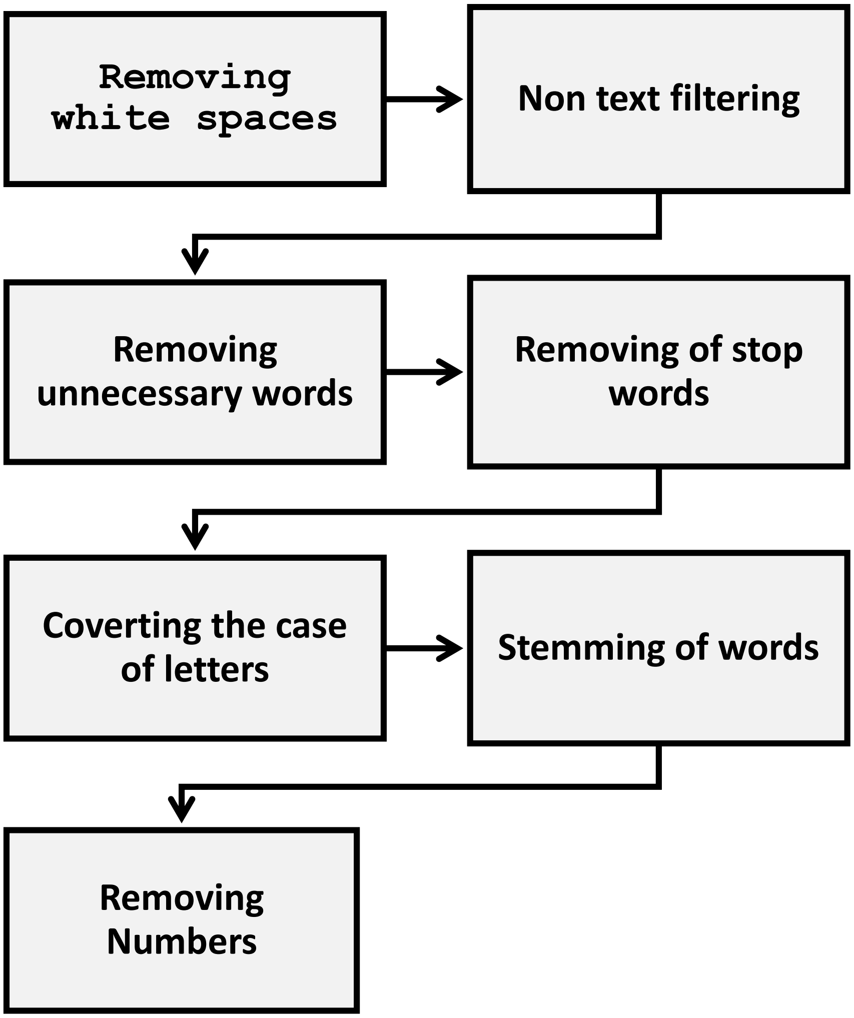


Figure 1: The Data Cleaning Unit that generates plain text for analysis

Unnecessary numeric values. Finally, we create a module as shown in figure 1; this module can be reusable for analysis of various different data sources.

The output of the preprocessing unit (Figure 1) is plain text, which can be used for data analysis. Typically, topics are identified by finding the special words that characterize documents about that topic. For instance, articles about soccer tend to have many occurrences of words like “football”, “ball”, “Mid-field”, “corner kick”, “foul”, “goal” etc. Once we have classified documents based on topics in this case soccer, it is not hard to notice words such as these frequently. However, until such classification is made, it is not possible to identify these words as characteristics. Thus, classification starts by looking at documents and finding significant words in those documents. We might initially guess significant words as ones that are more frequent, but when we see the results it would be the words like “the” or “and” also known as the most significant stop words, this can be eliminated by implementing the preprocessing steps again. The most important indicators are the less frequent words that are rare, but not in all cases. Sometimes words that are rare might mislead, in the context of soccer a word like “Pitch” is very rare but if we classify based on that, we would be in an undecidable state because it might be the same in a different game i.e. Cricket, Football. Thus, we must consider the combination of words across different documents and arrive at an opinion about the document. Based on this concept we have a document term matrix to help us get the frequent words. Information regarding diseases that occur among humans is considered as the foundation for biomedical research for identifying drug targets, understanding pathways relevant to novel treatment and combining clinical care and biomedical research (Warren et al, 2014). We need a standardized representation of human diseases to map across resources and support development of computational tools that would enable data analytics and integration. Perhaps the most common belief in qualitative research is that content analysis deals with finding the frequent words to arrive at the greatest concerns in the data set. While this notion holds correct for certain cases, there are several counterpoints to consider when using simple word frequency to explore problems of high significance. Kshitij et al, 2012, in their analysis focus on creating a computational framework for context specific integration of biomedical analysis based on free text literature.

Daniel et al, 2014 described the need of observing user groups behavior while creating information systems. Development of such a mechanism involves different types of users based on the familiarity with the domain, like novice, experts, specialist and so forth in accordance with their knowledge level in the field of study. These characteristics heavily influence user's behavior while interacting with the systems; these insights can be very useful when developing a system. Humans navigate an information repository (like Wikipedia) generally without the knowledge of network topology in its eternity. Stanley Milgram et al,(1960) in their small world experiment sent participants in Boston and Nebraska letters containing information about a target person and measured the length of the path (number of intermediate people which each one approaches) each one adopted before arriving at the target. The median length of the path has six intermediates famously coined as "six degrees of separation". The result illustrated in the small world phenomenon proved that to establish communication between two different persons in the United States there are only a few hops needed. The problem of merging ontologies with rules is vital in the semantic world. A straightforward combination of ontologies would result in an undecidable state, which might influence the analysis in multiple ways. To evade these complications a simple form of classification based on coupling rules and ontologies is introduced. They aim at bringing the description logic and their corresponding rules together with strict semantic integration and strict semantic separation the former is called loose coupling and the later known as tight coupling. These discoveries show that humans are very effective at finding short paths based on the local offline as well as online social networks.

Discovering links between syntactically and systematically similar terms across ontologies has been a traditional approach of integrating ontologies in the domain of biomedical engineering (Silva et al, 2003). These approaches relate terms with similar meanings but do not expose any relationships between apparently distinct functional spaces like disease, drugs and anatomy. Approaches like machine learning, natural language processing and graph matching also tend to focus on mapping synonyms across ontologies. Instance based methods that were proposed with modifications as a part of ontology alignment initiative generally catered to the traditional ontology alignment using synonyms. For the biomedical domain, the data is highly incongruent, modest ontological methods prove unworthy and the computational complexities that accompany with these datasets prove to be infeasible. Therefore, a context specific ontology is derived based on the dependent functional links between ontological concepts that occurs on free text literature. The prerequisite of such an analysis is a massive amount of unstructured data with co-occurring ontological terms, which serves as the basis of future analysis in this study.

CHAPTER 3: PROBLEM STUDIED IN THIS WORK

The focus of this study is on the social media impression on public health. Therefore, after surveying the literature on different aspects of data gathering, cleaning, analysis and visualization, we designed our experimental part of the study. In more detail, in this part we focus on the design of experiments, the methods of data collection, and the type of data analysis.

3.1. Design of the Experiments

The experiments are about collecting information from social media about disease outbreaks and analyzing the data to find specific outbreaks that exist in the corpus, the place of origin and compute the strengths of their co-occurrence. The experiments gradually begin with specific observation, which are used to produce generalized theories and conclusions drawn from the research. Our experimental study includes two parts, the data-gathering phase and the data analysis phase. The data-gathering phase is about experimenting with different methods of extracting data from web pages. This phase also has the data cleaning part used for cleansing data and creating a corpus with plain text. The data analysis phase involves exploration of different methods employed in the analysis of data to find out diseases that are prevalent across the corpus.

3.2. Data Gathering

In this phase of the experiment, data is gathered from various sources of social media like Facebook and Twitter. We focus on certain pages which concentrate on news and disease spread like CNN, WHO and Infectious Disease News[15] which are completely dedicated for disease surveillance. To maintain diversity of data across different domains, we created the repository focusing on sources that could provide earlier estimates of epidemic dynamics and sources where near in real-time data is available for consumption. According to Chunara et al, (2011) in his research on using social and news media to estimate epidemiological patterns in the 2010 Haitian cholera outbreak. It was proved that Informal data (Twitter and HealthMap16) combined with official data (government sources) in an outbreak setting helps in timely estimates of disease dynamics. As discussed earlier, information from various pages are populated to form a corpus (Figure 3) across different domains. During the initial phase of the project, we created a parser that could extract text from webpages using the python-based package Beautiful Soup[17]. The text yielded had html tags and URL`s, with further modifications to the code, a parser which extracted only data from web pages was implemented. Using csv package we were able to insert data into a .csv file. According to Marcus et al, 2014, in their research states the adoption rate of Facebook by news agencies is higher than Twitter. Analyzing the restrictions posted by Twitter in regards with the number of characters (140-character limit), Facebook is an ideal platform allowing users to express their feelings without any restrictions. This characteristic makes it an ideal platform for analytics as the burden for

[15] www.healio.com/infectious-disease

[16] www.healthmap.org/en/

[17] www.crummy.com/software/BeautifulSoup/bs4/doc/

grouping the posts for a particular issue is taken care by the internal mechanism of Facebook itself. Now, the focus shifted on extracting data from Facebook with the emphasis of extraction using the Graph API v2.8[18], this was possible with the RFacebook package in R. The new version of Graph API Facebook insisted the use of Open authentication (OAuth) mechanism. Traditional approach of password authentication is highly vulnerable to theft; OAuth secures the communication by a three-way handshake method. It also improves the confidence among the users, as the methodology does not need users to share their passwords across different third party applications.

The authentication of third party API like RFacebook is as follows:

1. Applications must be created in Facebook to access its data; this is done by creating an application in the Facebook developer's site. This process creates an application ID and an App secret which are used to register the application in Facebook.
2. Developers are required to authenticate with their credentials to view the App secret. In this case, we have to invoke the RFacebook package and feed the application id and app secret in its fb_oauth function and provide additional information regarding extended permissions. This ensures that public data can be seamlessly scraped from Facebook for analysis.
3. The Facebook API needs the URL on which the application is hosted for site validation. This is achieved by pasting the applications host URL to the default site URL on Facebook application settings.
4. R invokes the application URL page that was entered in the previous step in the default browser and authenticates the application for data extraction.

[18] developers.facebook.com/blog/post/2016/10/05/graph-api-2.8/

The application id and the app secret do not change and can be cached by the application for future use. The process is summarized in figure 2, which explains the handshake carried out between different modules to establish authentication. We were able to extract data based on the page information provided in the query along with date time stamps.

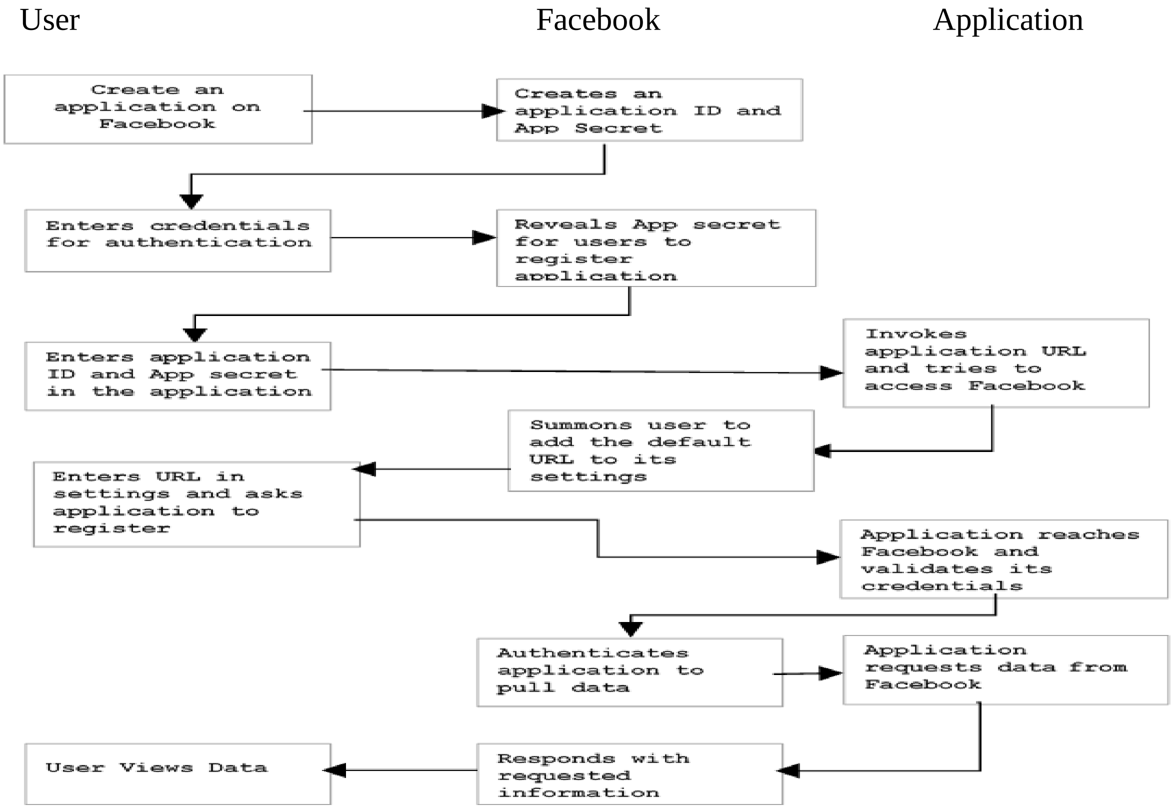

User
Facebook
Application
Create an application on Facebook
Creates an application ID and App Secret
Enters credentials for authentication
Reveals App secret for users to register application
Enters application ID and App secret in the application
Invokes application URL and tries to access Facebook
Summons user to add the default URL to its settings
Enters URL in settings and asks application to register
Application reaches Facebook and validates its credentials
Authenticates application to pull data
Application requests data from Facebook
User Views Data
Responds with requested information

Figure 2: Steps for accessing data from Facebook

Data collected in this mechanism is combined to form a corpus which can be used for data processing. Figure 3 describes the composition of the corpus and the different sources that were a part of it. Posts are a form of raw data with huge portions of noise in the form of punctuations, repetitions, stop words and white spaces. The next phase deals with data cleaning to implement processes to clean data and prepare it for further analysis.

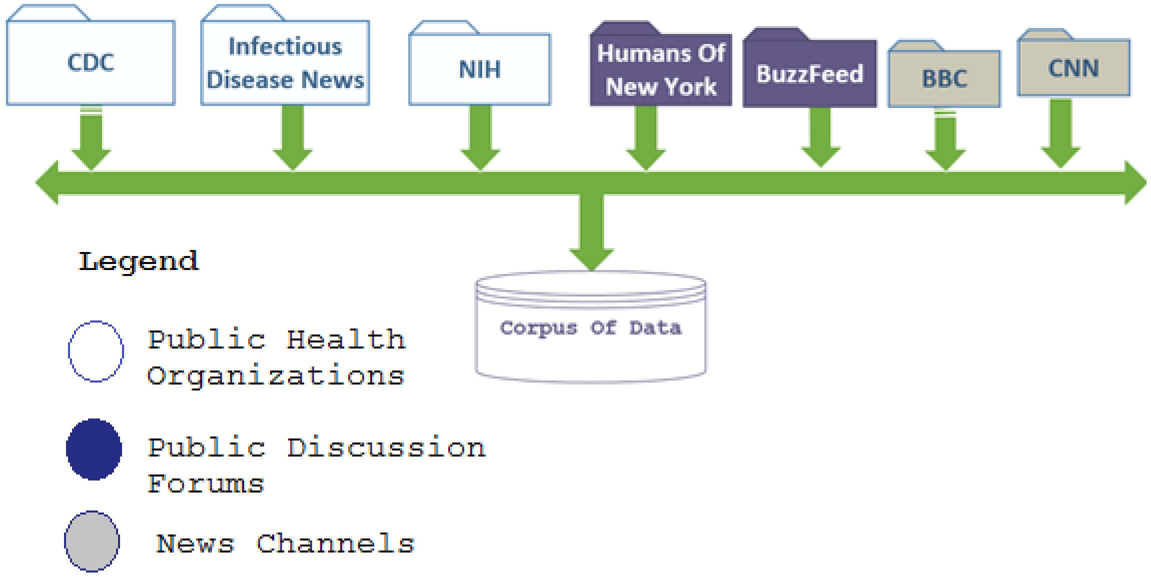


Figure 3: Data from various different sources forming a corpus

To maintain uniformity across the corpus, letters were converted to lower cases, thus avoiding errors due to case differences. Stop words are a set of commonly used words in any language, not just English. The significance why stop words are serious to many applications is that, if we eliminate some words that are commonly used in any given language to add a grammatical meaning to a sentence we can concentrate on the important words. For example, in the context of a search engine, if a query “how to create a search engine” is issued, if the search engine tries to search for pages that contain “how”, “to”,” a”,” create”,” search” and “engine”. It will return all those pages that match their cases, but if they search for pages like “how”,” a” and “to” there would be many pages because these words are very commonly used in English language. If these terms are disregarded the actual focus on retrieving pages that contain “create”,” search” and “engine” can be provided, which makes the process of search and retrieval faster. This also brings in more pages that actually match the context of search. To accomplish this task, we created a minimal stop word list which encompassed determiners that are usually followed by nouns like the, a, an, and so forth, coordinating conjunctions which act as a means to connect words, phrases and clauses and prepositions that express temporal or spatial relations like in, under, towards and before were inducted. Since we used a domain specific analysis we had to consider a special set of stop words like “mcg”, “dr” and “patient” as they have less discriminating power in building intelligent analytics compared to terms such as “heart”, “failure” and “diabetics”. This prompted us to create a domain specific stop words list as opposed to using a generic tool based list of words.

Stemming is one of the first steps in the information retrieval pipeline (Salton, 1971). A stemmer or stemming algorithm aims at obtaining the stem of a word, i.e. its morphological root, by cleaning the affixes that carry grammatical or lexical information about the words. The affixes do not modify the concept the words are related, as the semantic informality has been proven in the literature, especially in languages that are highly inflective (Popovic and Willett, 1992) and in short documents (Krovetz,1993), in terms of recall and precision. A stemming algorithm has three main purposes. The first one consists of clustering the words according to the topic. Many words are derivations from a same stem and they belong to the same concept (e.g., drive, driven, driver). These are generated through appended affixes but, in general and more specifically in English, only suffixes are considered, as generally prefixes and infixes modify the word and stripping them would lead to errors of bad topic determination (Hull, 1996). Documents pertaining to certain topics like medicine or biology suffixes maintain the concept of word. Among these suffixes two types of derivations are considered (Krovetz, 1993). Inflectional derivations related to gender, number, case, mood or tense. Stripping these words do not actually provoke a change in the part- of-speech of the original language nor its meaning. On the contrast removing derivational suffixes that deal with the creation of new words based on existing words with which it shares meaning (example, words ending with -ize, -ation, -ship etc) allows its stem to be obtained, which is nearly its morphological root; this process can identify thematically related words by matching their stems. The second step carried out by a stemmer or a stemming algorithm is connected to the information retrieval process, where the stem of the word can be used to index the document based on their topics, as their terms are grouped by stems. This improves the process of information retrieval and helps to fetch information with least number of searches. The final step in stemming is conflation of words sharing the same stem, this leads to the reduction of the dictionary, and the space needed to store the structures used during information retrieval. Following these steps, we were able to create a list of words to be considered as roots. This helped us in reducing the errors due to analysis of words that exhibit the same meaning.

3.3. Data Analysis

At the end of the data cleaning phase, data collected from various sources are combined together into a single repository. The initial task was to mine for the details regarding diseases that were present across the corpus. For this purpose, we created the document term matrix. This matrix provides us the number of occurrence of a word in the document. The sum of occurrence of individual terms was considered across different files in the corpus and a table with words and their frequencies was created. As a part of research called DOID[19], ontologies of diseases are provided by aggregating disease names from different source as a .csv file we used this as a legend to compare and find out the diseases reported in the corpus using SQLDF package in R. A SQL query is written to compare data and the result lists only components that are present in both the tables. The final data set obtained consists of diseases and their frequencies. This result helps us to get information about the different diseases that are present in the repository. Since the query uses a huge amount of memory and do not measure any relationship between the words we can use co-occurrence to compute the occurrence of diseases, the frequency is a measure without any relationships. The co-occurrence is a relative measure of strength between two words, we use this measure to calculate the support and confidence across the repository.

Analyzing the repository to find disease outbreaks and their co-occurrences indirectly reveals the relationships between diseases and their corresponding location, and serve as an indicator for governments to take precautionary actions. The following analysis is based on QDA miner[20], a software used for text analytics. In the initial phase, we use the associative rule to find co–occurrences of diseases, and then we can use the measure of confidence and support to calculate the probability with which the disease can be predicted with confidence.

3.3.1. Observation 1

When analyzing data using QDA miner for the co-occurrence of words using the associative distance metric the below distribution was observed, the ontological

[19] disease-ontology.org/ where the disease ontologies are provided here as a .csv file

[20] provalisresearch.com/products/qualitative-data-analysis-software/

representation in figure 4 is the analysis of data from the period between Jan2013 to Aug 2016.

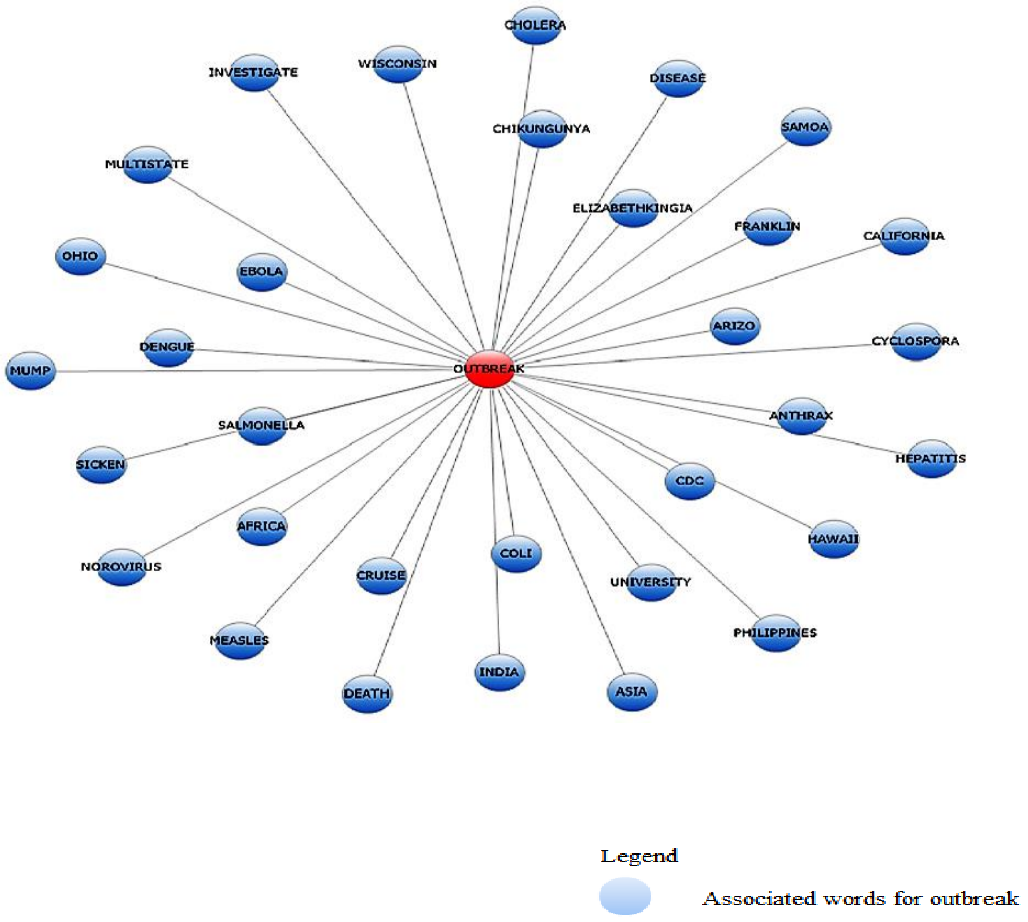


Figure 4: Outbreaks and their location

Figure 4 represents the words associated with the word outbreak; the circles represent the words that are a part of posts by users. The posts are then separated for words and arranged based on the frequency of their occurrence. These words are further mapped based on their association established in the posts. Posts include the areas where the diseases originated as well as various diseases that has been reported. The high-level knowledge of different diseases and outbreaks help us to channelize our focus on a particular place or disease. For instance, when posts related to Philippines are analyzed from another corpus of data formed by combining posts it shows that there are outbreaks of diseases like measles and rabies as shown in figure 4.

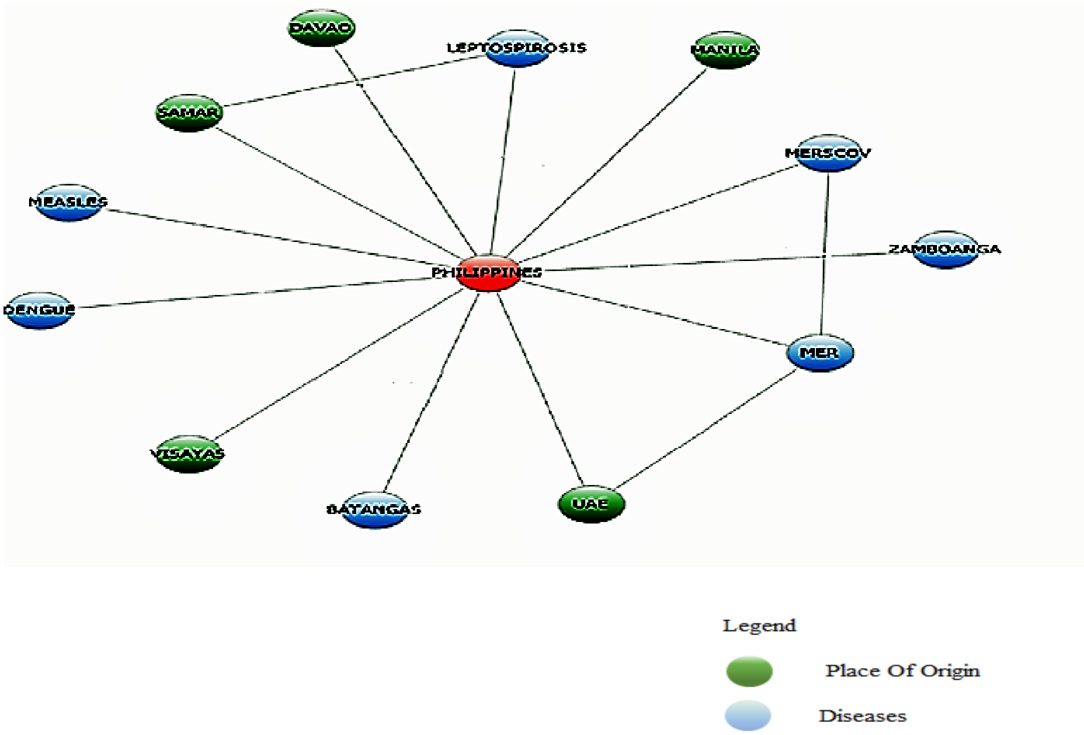
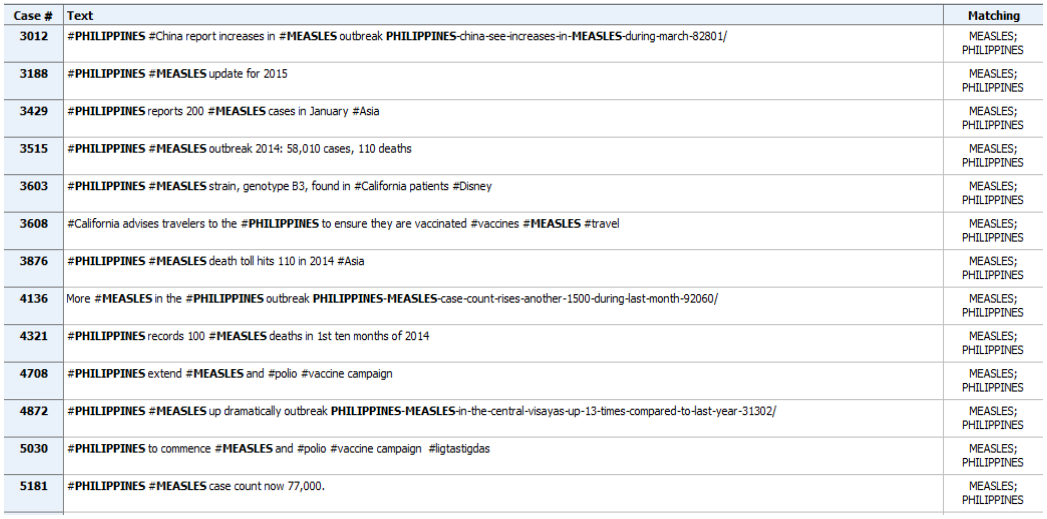


Figure 4: Diseases and their location at Philippines

For further analysis, we explore the case of measles outbreak reported in Philippines in the following posts:



Case #	Text	Matching
3012	#**PHILIPPINES** #China report increases in #**MEASLES** outbreak **PHILIPPINES**-china-see-increases-in-**MEASLES**-during-march-82801/	MEASLES; PHILIPPINES
3188	#**PHILIPPINES** #**MEASLES** update for 2015	MEASLES; PHILIPPINES
3429	#**PHILIPPINES** reports 200 #**MEASLES** cases in January #Asia	MEASLES; PHILIPPINES
3515	#**PHILIPPINES** #**MEASLES** outbreak 2014: 58,010 cases, 110 deaths	MEASLES; PHILIPPINES
3603	#**PHILIPPINES** #**MEASLES** strain, genotype B3, found in #California patients #Disney	MEASLES; PHILIPPINES
3608	#California advises travelers to the #**PHILIPPINES** to ensure they are vaccinated #vaccines #**MEASLES** #travel	MEASLES; PHILIPPINES
3876	#**PHILIPPINES** #**MEASLES** death toll hits 110 in 2014 #Asia	MEASLES; PHILIPPINES
4136	More #**MEASLES** in the #**PHILIPPINES** outbreak **PHILIPPINES**-**MEASLES**-case-count-rises-another-1500-during-last-month-92060/	MEASLES; PHILIPPINES
4321	#**PHILIPPINES** records 100 #**MEASLES** deaths in 1st ten months of 2014	MEASLES; PHILIPPINES
4708	#**PHILIPPINES** extend #**MEASLES** and #polio #vaccine campaign	MEASLES; PHILIPPINES
4872	#**PHILIPPINES** #**MEASLES** up dramatically outbreak **PHILIPPINES**-**MEASLES**-in-the-central-visayas-up-13-times-compared-to-last-year-31302/	MEASLES; PHILIPPINES
5030	#**PHILIPPINES** to commence #**MEASLES** and #polio #vaccine campaign #ligtastigdas	MEASLES; PHILIPPINES
5181	#**PHILIPPINES** #**MEASLES** case count now 77,000.	MEASLES; PHILIPPINES

Table 1: Posts on Measles and their location at Philippines

From the information above, we can get the support and confidence of Philippines having an outbreak the confidenceof an outbreak that is reported in Philippines is represented as

$Confidence\,\{\,Philippines \rightarrow Outbreak\}$

$$Confidence\,\{Philippines \rightarrow Outbreak\} = \frac{Support\{Outbreak\ \cap\ Philippines\}}{Support\{Philippines\}}$$

The number of occurrences of Philippines in the corpus is represented as *Support{Philippines} = 153*

The number of Co-occurrences of Outbreak and Philippines is represented as

$Support\{\,Outbreak\ \cap Philippines\} = 32$

The confidence is calculated as mentioned in the formula above

$Confidence\{\,Philippines \rightarrow Outbreak\} =\ 20.91$

It can be proved that in the repository when the support that Philippines has an outbreak is around 48 and has a confidence interval of 20.91 percent. The confidence is low due to the vastness of the repository. Moving on to the outbreak of Measles, we need to find the support of measles when browsed for posts in connection with Philippines. This can be calculated similarly as mentioned in the previous observation.

$$Confidence\{\,Measles \rightarrow Philippines\} = \frac{Support\{Measles\ \cap Philippines\}}{Support\{\,Measles\}}$$

From the observation above it can be said that Philippines has measles can be predicted with 16.34 percent confidence. This method can be applied to different diseases and we can compute the support and confidence across the repositories for its occurrence.

3.3.2. Observation 2

Zika Virus vs Flu: In this analysis, we try to find the association exhibited by the diseases. Figure 5 shows the close relationship between the spread of Zika virus and various factors responsible for its spread. This association helps us to understand different locations where the disease is spread in the corpus. For example, there is a strong association between Zika Virus and Miami compared to Brazil and Olympics. When analyzing the posts, the number of reports on Zika at Miami was greater than Brazil, which is a clear indicator that awareness of Zika is high in Miami compared to Brazil.

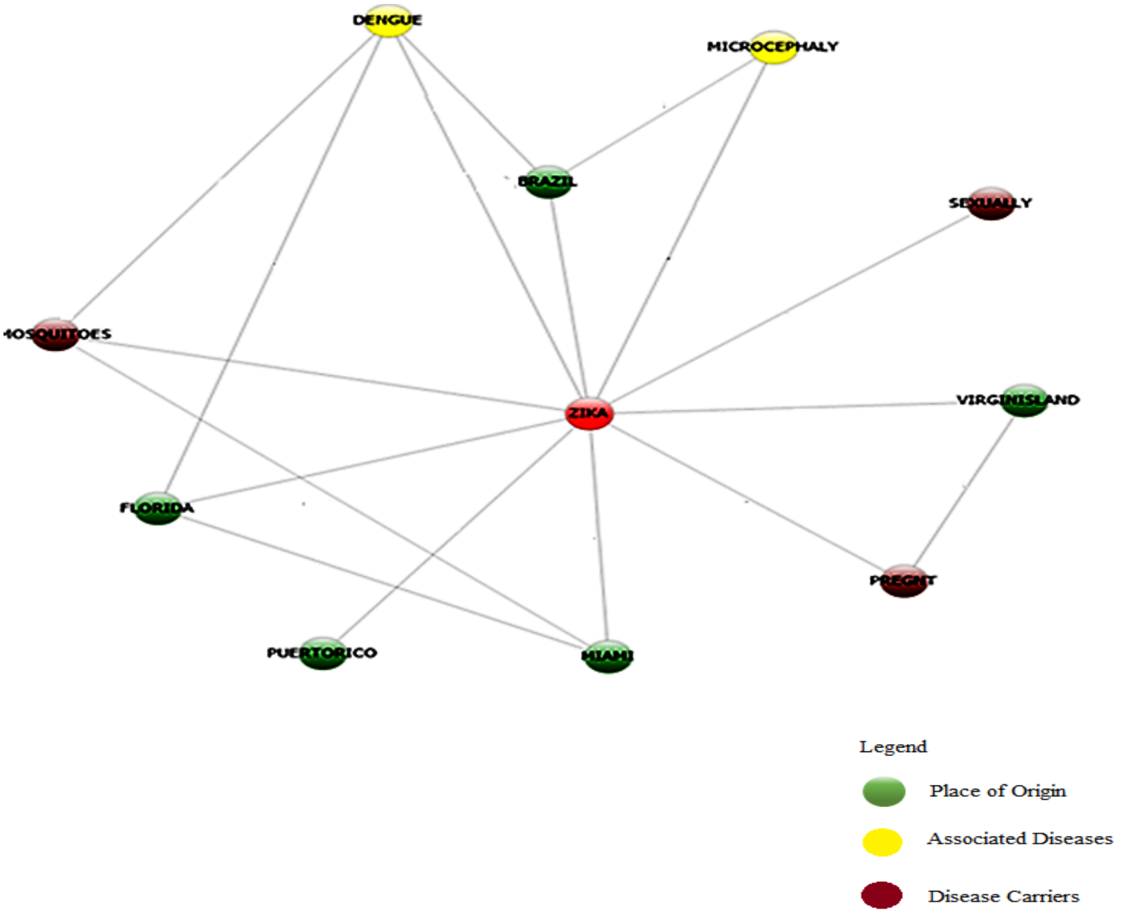
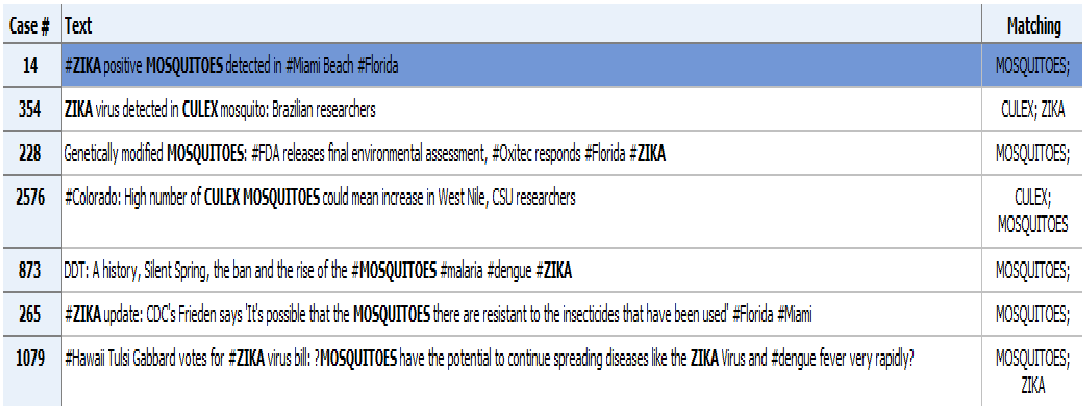


Figure 5: Visualization of Zika outbreak

On Further, analysis there is a strong relationship between Miami and DADE County. Ontological representation sometimes specifies the carrier of diseases, for example in this case, we can see mosquitoes, sexual contacts, and through pregnancy as the major reasons behind diseases even though the number of posts relating to the carriers are low, these posts can help us get new insights of the evolution of diseases and its viruses.



Case #	Text	Matching
14	**#ZIKA** positive **MOSQUITOES** detected in #Miami Beach #Florida	MOSQUITOES;
354	**ZIKA** virus detected in **CULEX** mosquito: Brazilian researchers	CULEX; ZIKA
228	Genetically modified **MOSQUITOES**: #FDA releases final environmental assessment, #Oxitec responds #Florida **#ZIKA**	MOSQUITOES;
2576	#Colorado: High number of **CULEX MOSQUITOES** could mean increase in West Nile, CSU researchers	CULEX; MOSQUITOES
873	DDT: A history, Silent Spring, the ban and the rise of the **#MOSQUITOES** #malaria #dengue **#ZIKA**	MOSQUITOES;
265	**#ZIKA** update: CDC's Frieden says 'It's possible that the **MOSQUITOES** there are resistant to the insecticides that have been used' #Florida #Miami	MOSQUITOES;
1079	#Hawaii Tulsi Gabbard votes for **#ZIKA** virus bill: ?**MOSQUITOES** have the potential to continue spreading diseases like the **ZIKA** Virus and #dengue fever very rapidly?	MOSQUITOES; ZIKA

Table 2: Zika related posts

Table 2 illustrates the relationship between Zika and different locations the disease is spread from the collected posts. It is important to know that sometimes some relationships can be found which are solely based on research proposals and/or research reports. Such information can be misleading. Therefore, we need to scrutinize the information for validity by a knowledgeable resource (supervised learning) to ensure proper information is obtained for the analysis. For example, we can observe one instance of co –occurrence between Culex and Zika[21] which reports Culex is now a carrier of Zika virus. This information was a proposal provided by researchers in Brazil but that's not

[21] www.sciencedaily.com/releases/2016/09/160922104408.htm

proved to be genuine. Hence, we emphasize the need for a domain expert to guide us to make proper decisions in verifying the results of such analysis.

Flu, a seasonal disease is depicted with similar way with a little focus on the variants of disease, it can be seen that different types of flu is reported, like Avian, influenza and symptoms like cold, cough and sneeze. This also provides an important information relating to the source of spreading these diseases, example bird-flu is spreading in Minnesota from Turkey flock also they were responsible for spreading AVIAN and Influenza. There is a clear association that bird flu is spreading in Egypt and Jiangxi (China).

Observing both scenarios it is revealed that Zika virus has spread across different parts of the world where as flu is mostly prevalent in Asia and Africa. It also specifies the different researches that is going on around the globe for diseases. This might set an alarm for health agencies all over the globe to take precautions against diseases to stop them before it becomes an epidemic. Even though, the support and confidence intervals are low, these markers should be included in the overall analysis as they serve as indicators for disease spread.

3.3.3. Observation 3

An interesting observations made during this study is the analysis of outbreak of diseases across different universities. Figure 6 provides information regarding various diseases outbreaks and the researchers conducted across different universities. This information can be used to analyze the relationship between diseases and their outbreak and various researches that are conducted across different universities.

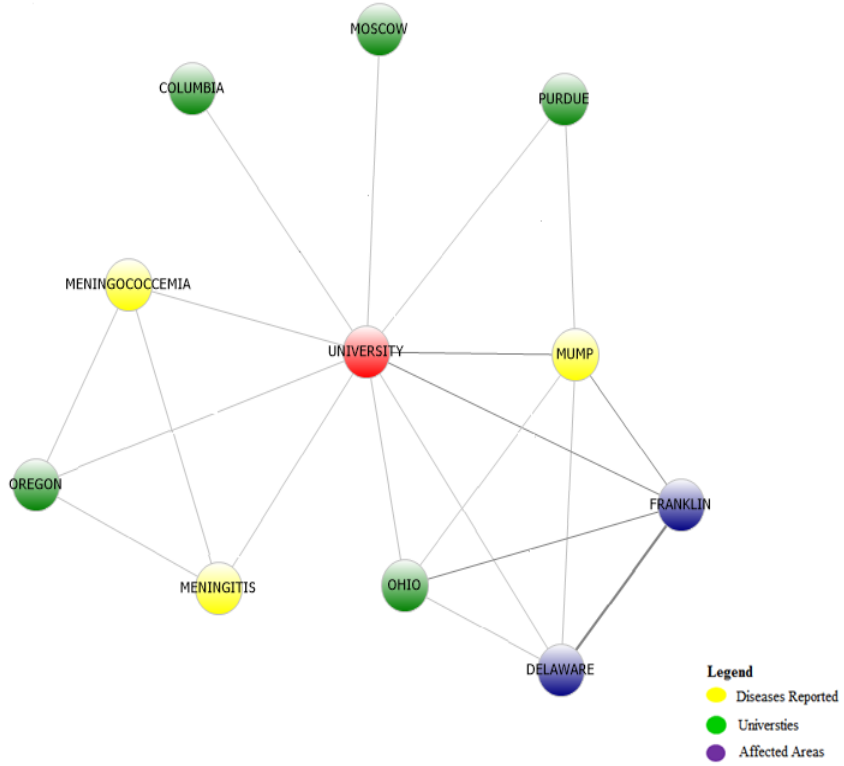


Figure 6: Visualization of Diseases reported in universities

From figure 6, it is observed that outbreaks of Mumps and Meningitis have been reported across different universities and they have spread to locations like Delaware and Franklin. This association is a strong indicator of outbreak and spread of specific diseases. Data from certain universities revealed the ongoing researches, which clearly proves that social media is a way of dissipating information about diseases. Reliability is a major factor that should be taken into account, the example above, has information about universities with ongoing research and the universities where an outbreak has been reported. In general, users perceive there are outbreaks reported across all these universities, but when we retrieve posts concerning each and every association the information about researches conducted for certain disease are revealed. This fact proves that results obtained by analysis from social media should undergo expert analysis before release. .

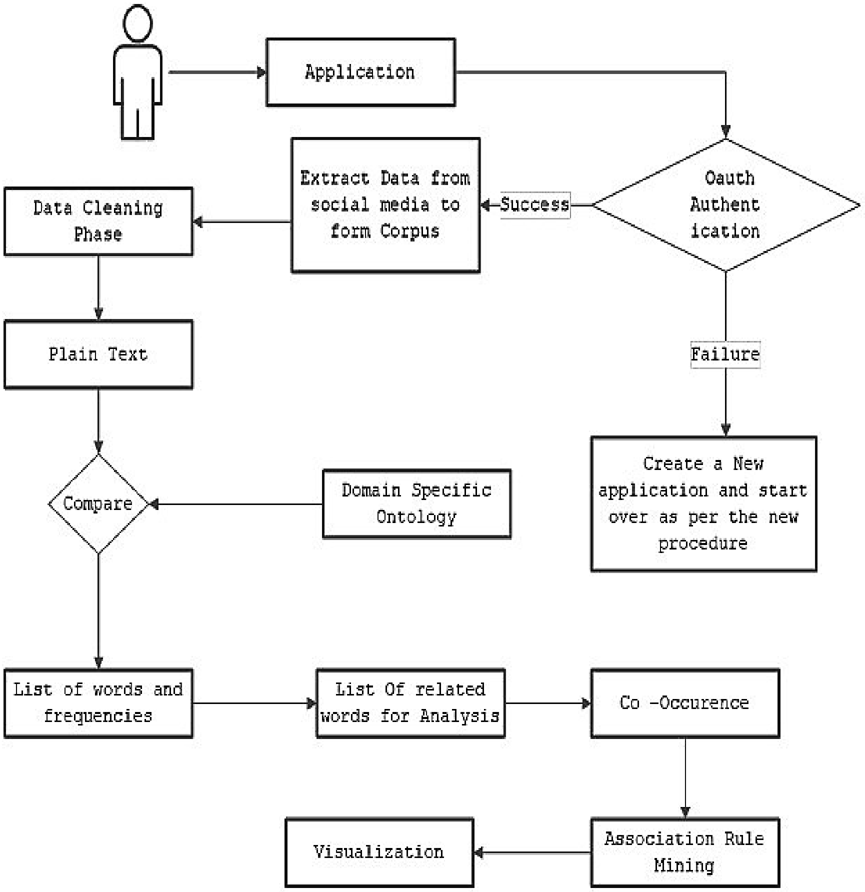

Application
Oauth Authent ication
Success
Extract Data from social media to form Corpus
Data Cleaning Phase
Plain Text
Failure
Create a New application and start over as per the new procedure
Compare
Domain Specific Ontology
List of words and frequencies
List Of related words for Analysis
Co -Occurence
Association Rule Mining
Visualization

Figure 7: Workflow for the steps to be performed during analysis from social media

From the analysis above we can establish a procedure for analyzing the informal information from social media. The steps at a high level can be classified starting with a data gathering phase for extraction of data from different social media followed by data cleaning responsible for removing unwanted noise in the data and finally the data analysis phase responsible for mining and visualization of results. This procedure as in figure 7 can be extended to different domains for effective analysis of Social media.

CHAPTER 4: CONCLUSION AND FUTURE ENHANCEMENTS

In this study, we have shown how social media resource Facebook can be utilized to detect valuable information regarding diseases and outbreaks. More specifically, using the posts from users across the globe, we constructed an ontology to find the associations that existed between certain diseases and the place of occurrence. Furthermore, we showed how such associations could be used to compare outbreaks of certain diseases and measure the impact across the globe. We also analyzed certain disease outbreaks in different universities and demonstrated how Facebook is used as a means to communicate about diseases across higher education.

As discussed, in this study we focused on the problem of detecting disease outbreaks and measured the support and confidence using the associations and co – occurrences of certain pairs of words. This method can be further extended to find the origin and the rate of spread of each disease. This can go further beyond public health data and can be extended across diverse set of applications and domains such as distributed supply chain, e-commerce and even presidential election outcomes. At a deeper level, this work shows how social media expresses a collective wisdom (Asur, et al., 2010) which, when properly harnessed, can yield extremely powerful and close to accurate indicators.

REFERENCE

Popovic, M. & Willett, P. (1992). The effectiveness of stemming for natural-language access to Slovene textual data. Journal of the American Society for Information Science, 43(5), 384-390.

Osterrieder, A. (2013). The value and use of social media as communication tool in the plant sciences. Plant methods, 9(1), and one.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., & Kibbe, W. A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic acids research, 40(D1), D940-D946.

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. The American Journal of Tropical Medicine and Hygiene, 86(1), 39-45.

Cleaton, J. M., Viboud, C., Simonsen, L., Hurtado, A. M., & Chowell, G. (2015). Characterizing Ebola Transmission Patterns Based on Internet News Reports. Clin Infect Dis. Clinical Infectious Diseases, 62(1), 24-31. doi:10.1093/cid/civ74

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Schriml, L. M. (2014). Disease Ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Research, 43(D1). doi:10.1093/nar/gku1011

Harris, A. L., & Rea, A. (2009). Web 2.0 and virtual world technologies: a growing impact on is education. Journal of Information Systems Education, 20(2), 137–144

Barton, H. and Grant, M. (2006) A health map for the local human habitat. The Journal for the Royal Society for the Promotion of Health, 126 (6). pp. 252-253. ISSN 1466-4240 Available from: www.eprints.uwe.ac.uk/7863

Zhou, X., Han, H., Chankai, I., Prestrud, A., &amp; Brooks, A. (2006, April). Approaches to text mining for clinical medical records. In Proceedings of the 2006 ACM symposium on Applied computing (pp. 235-239). ACM.

Yepes, A. J., MacKinlay, A., &amp; Han, B. (2015). Investigating Public Health Surveillance Using Twitter. ACL-IJCNLP 2015, 164.

Ananiadou, S., Kell, D. B., & Tsujii, J. I. (2006). Text mining and its potential applications in systems biology. Trends in biotechnology, 24(12), 571-579.

Oh, H. J., Lauckner, C., Boehmer, J., Fewins-Bliss, R., & Li, K. (2013). Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. Computers in Human Behavior, 29(5), 2072-2080. DOI: 10.1016/j.chb.2013.04.017

Lampos, V., &amp; Cristianini, N. (2010, June). Tracking the flu pandemic by monitoring the social web. In Cognitive Information Processing (CIP), 2010 second International Workshop on (pp. 411-416). IEEE.

Lerman, K., & Ghosh, R. (2010). . In International AAAI Conference on Web and Social Media.

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.

Prieto, V. M., Matos, S., Álvarez, M., Cacheda, F., &amp; Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. PloS one, 9(1).

Oh, S., Yi, Y. J., &amp; Worrall, A. (2012). Quality of health answers in social Q & A. Proceedings of the American Society for Information Science and Technology, 49(1), 1-6.

Laender, A. H., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. ACM Sigmod Record, 31(2), 84-93.

Milgram, Stanley. “The Small World Problem.” Psychology Today. Two, pp 60-67, 1967.

Milgram, Stanley. The Individual in a Social World: Essays and Experiments. McGraw-Hill, Inc., 1992.

Silva, N., & Rocha, J. (2003). Complex semantic web ontology mapping. Web Intelligence and Agent Systems: An International Journal, 1(three, 4), 235-248.

Strohmaier, M., Walk, S., Pöschko, J., Lamprecht, D., Tudorache, T., Nyulas, C., & Noy, N. F. (2013). How ontologies are made: Studying the hidden social dynamics behind collaborative ontology engineering projects. Web Semantics: Science, Services and Agents on the World Wide Web, 20, 18-34.

Lamprecht, D., Strohmaier, M., Helic, D., Nyulas, C., Tudorache, T., Noy, N. F., & Musen, M. A. (2015). Using ontologies to model human navigation behavior in information networks: A study based on Wikipedia. Semantic web, 6(4), 403-422.

Pirolli, P., & Card, S. (1999). Information foraging. Psychological review, 106(4), 643.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

Krovetz, R. (1993). Viewing morphology as an inference process. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 191-202. New York, NY: ACM Press.

Salton, G. (1971). The SMART retrieval system - experiments in automatic document processing. Upper Saddle River, NJ: Prentice-Hall, Inc.

Matheus, C. J., & Chan, P. K. (1993). Systems for Knowledge Discovery in Databases.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using internet searches for influenza surveillance. Clinical infectious diseases, 47(11), 1443-1448.

Keller, M., Freifeld, C. C., & Brownstein, J. S. (2009). Automated vocabulary discovery for geo-parsing online epidemic intelligence. BMC bioinformatics,10(1), 385.

Moynihan, R., Bero, L., Ross-Degnan, D., Henry, D., Lee, K., Watkins, J., ... & Soumerai, S. B. (2000). Coverage by the news media of the benefits and risks of medications. New England Journal of Medicine, 342(22), 1645-1650.

Chan, E. H., Brewer, T. F., Madoff, L. C., Pollack, M. P., Sonricker, A. L., Keller, M., ... & Brownstein, J. S. (2010). Global capacity for emerging infectious disease detection. Proceedings of the National Academy of Sciences, 107(50), 21701-21706.

Conway, M., Doan, S., Kawazoe, A., & Collier, N. (2009). Classifying disease outbreak reports using n-grams and semantic features. International journal of medical informatics, 78(12), e47-e58.

Riga, M., & Karatzas, K. (2014, June). Investigating the relationship between social media content and real-time observations for urban air quality and public health. In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) (p. 59). ACM.

Sokolova, M., Jafer, Y., & Schramm, D. (2012, September). Text Mining for Personal Health Information on Twitter. In Proceedings of the 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (p. 112). IEEE Computer Society

Paul, M. J., & Dredze, M. (2011, July). You are what you Tweet: Analyzing Twitter for public health. In ICWSM (pp. 265-272).

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Weeg, C. (2015). Psychological language on twitter predicts county-level heart disease mortality. Psychological science, 26(2), 159-169.

Kim, J. D., & Wang, Y. (2012, June). PubAnnotation: a persistent and sharable corpus and annotation repository. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (pp. 202-205). Association for Computational Linguistics.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Weeg, C. (2015). Psychological language on twitter predicts county-level heart disease mortality. Psychological science, 26(2), 159-169.