Adapting CRISP-DM Process for Social Network Analytics: Application to Healthcare

Full paper

Daniel Adomako Asamoah
Wright State University
daniel.asamoah@wright.edu

Ramesh Sharda
Oklahoma State University
ramesh.sharda@okstate.edu

Abstract

One of the key limitations about research involving big data is the lack of a sound methodological process that drives the conceptual and analytical questions posed to the data. In this study, we adapt the popular CRISP-DM process to analyze large volumes of unstructured data to generate analytical insights. We add specificity to the CRISP-DM methodology. Specifically, we propose "Cross Industry Standard Process for Electronic Social Network Platforms (CRISP-eSNeP)", as an extension to the CRISP-DM methodology. Our methodology focuses on efficient pre-processing of large and unstructured electronic social network data. We illustrate our arguments by applying this methodology to understand the relationship between user influence and information characteristics as depicted on the Twitter microblogging platform.

Keywords

Big data, analytics, methodology, healthcare, Major Depressive Disorder (MDD), CRISP-eSNeP, CRISP-DM, social networks.

Introduction

"Big Data is not about the data". Gary King of Harvard University made this comment while making the point that the real value of collecting large volumes of data is in the extent to which interesting knowledge can be extracted from the data. While it is much touted that Big Data is the new gold, we believe that Big Data is rather the new gold ore. Data alone does not guarantee access to actionable insights. Rather, value is created when in-depth insights are derived from the data (gold ore) to enable one to generate meaningful knowledge (refined gold).

Data mining processes such as the Cross Industry Standard Process for Data Mining (CRISP-DM) have been developed as a guideline for data mining projects (Shearer 2000). However, such processes, developed prior to the 'data boom' age are without due cognizance to the amount and multi-structured nature of data generated by modern information systems. Methodological processes such as CRISP-DM could benefit from additional guidelines for generating insights from large volumes of electronic social network (SN) data.

Based on our experiences in working with large data sets, we contribute to analytics methods by adding specificity to the CRISP-DM methodology. Particularly, we propose the Cross Industry Standard Process for Electronic Social Network Platforms (CRISP-eSNeP) as an extension to the CRISP-DM methodology. Our extension focuses on efficient pre-processing, management and analysis of large and unstructured SN data. Based on the design science research method (Hevner et al. 2004), we generate a set of guidelines for assessing large volumes of electronic SN data.

A key contribution of this paper is the development of a methodological process CRISP-eSNeP, for managing and analyzing big data sets on electronic SNs, particularly in social science research. We develop this process through lessons learned while implementing this methodology in a healthcare domain, where we focus on patient self-management of Major Depressive Disorder (MDD). We utilize our process to study the relationship between an electronic SN structure and the type of information disseminated on such networks.

In the next section, we discuss related studies about big data implementations and methodologies. Next, we briefly describe the Apache Hadoop platform used in this study and subsequently present an introduction of the disease of interest. Thereafter, we present our data analytics methodological process (CRISP-eSNeP). This section also presents our argument for the use of this process through analysis of a Twitter data set. We conclude with a set of Guidelines for utilizing CRISP-eSNeP for analyzing big data from SNs. We also present suggestions on ways our methodological process can be extended to other disciplines and other large data sources.

Literature Review

Sagiroglu and Sinanc's (2013) definition of big data encompasses the basic components of general definitions of big data. They define big data as a term for "massive data sets having large, more varied and complex structure with difficulties of storing, analyzing and visualizing for further processes or results". Such difficulty in storing and analyzing the data could also be due to velocity of data acquisition, data variability and complexity.

Electronic SNs have opened new frontiers to big data sets for social science research. Even though new and advanced data management systems such as Apache Hadoop help social scientists easily access large volumes of data and generate deeper insights, there still persists the hurdle of a sound methodological process that can streamline the effective management and analysis of such big data.

Furthermore, in social science research, large volumes of potential data from SNs pose a challenge to statistical testing methods, since such research has traditionally relied on relatively small samples of data using established methodologies. On the one hand, SNs provide a plethora of data. On the other hand, large sample sizes pose a challenge to statistical significance testing (Dubitzky et al. 2007).

Several analytical approaches have been utilized in the past. Most studies have performed analysis specific to a particular project only. For instance, Kwak et al. (2010), crawled the entire "Twittersphere", resulting in the extraction of information on 41.7 million user profiles and 1.47 billion relationships. Whereas their data extraction and analysis approach ensures access to a significant depth of insights, their approach lend itself to a more effective analysis of the structural characteristic of nodes and the relationships between them as depicted on Twitter rather than the generation of direct knowledge and insights based on both structure of the social graph and the content of posted messages.

Some studies have also extracted limited amount of messages in a defined social graph for analysis. Shi et al. (2014) extracted 65 root tweets over a 140-day period. In addition, they tracked and recorded retweet activities for each root tweet for the next five days. This approach presents a more focused study, albeit at the expense of a more generalizable and broader spectrum of analysis and subsequent insight that can be generated.

Despite the plethora of analytical approaches for big data, less has been done on actual methodologies that can support the conceptualization of relevant research questions. Previous studies mostly focus on methodological challenges in working with big data and prescribe options to address such challenges (Gandomi and Haider 2015). For instance, methods utilizing video analytics (Hu et al. 2011), SN analysis (Heidemann et al. 2012) and predictive analytics, have been proposed for studying and extracting insightful information from data.

CRISP-DM has emerged as a de facto mechanism for data analytics (Shearer 2000; Wirth 2000). Whereas the approach is generalizable for most forms of data analytics, it does not adequately respond to more recent research hurdles posed by large volumes of data. Although some recent studies such as the development of the HACE theorem to model big data characteristics (Wu et al. 2014) have been produced, such studies are deficient in providing a generalizable and yet parsimonious approach to managing complex data patterns as generated on electronic SNs. Tinati et al. (2014) notes that in social research, methodological hurdles to managing huge amount of data puts a cap on the extent to which ontological and epistemological questions can be asked. In this study we recognize the lack of a sound methodology that drives the conceptual and analytical questions posed to large volumes of data extracted from electronic SNs. This begs the question, "what is a concise methodological process that will leverage large volumes of social network data in social science research?"

Against this background we propose CRISP-eSNeP, an extension to the CRISP-DM methodology, as a guideline to manage, analyze and generate insights from large volumes of data acquired from electronic SNs.

Big Data Platform

The Hadoop big data platform is an emerging architecture in data analytics and is used as a tool in this study to derive underlying insights about disease self-management information provided on Twitter.

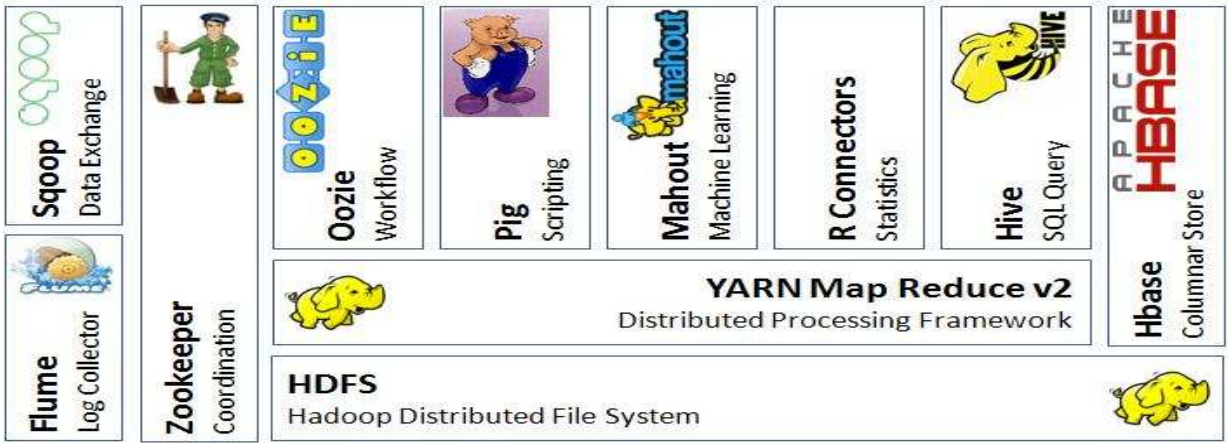

Figure 1. Apache Hadoop Ecosystem (Adapted from (Singh 2014))

The platform usually works in concert with existing data mining and business analytics tools (figure 1). Key components of the Apache Hadoop framework (Apache Software Foundation 2014) are Hadoop Common which consists of common utilities that support other modes on Hadoop, a Hadoop Distributed File System (HDFS) which is a fault tolerant distributed file system that provides fast throughput to data, a Hadoop YARN which is a framework for job scheduling and cluster management and MapReduce which is a system for processing data in parallel (Dean and Ghemawat 2008). Some related projects on this open-source platform are Apache Hive and Apache HBase. For instance, Apache Hive is a data warehouse infrastructure that mimics an SQL database and supports instant data querying.

Role and Mechanics of Twitter

Research utilizing data from Twitter is increasing exponentially as researchers explore how relationships formed using electronic SNs can help understand human and organization behavior. Also unlike other SNs, its use as a micro-blogging platform encourages users to maintain public profiles, hence allowing access to large amount of research data.

Twitter gives patients more avenues and control in self-managing their diseases using data disseminated by others. It also fosters increased participation and direct user engagement (Hesse et al. 2011) through the concept of the 'Power-of-Many", where knowledge wielded by several non-experts on a subject matter is considered to be at par with that of any single expert.

The mechanics of communication on Twitter occurs in a directed graph where relations are not necessarily reciprocal. A user can have either followers or followees. A tweet posted by a user is shared by only his followers and not his followees. A message sent in response to an earlier tweet is called a 'retweet'. Conventional means of commenting or replying to earlier tweets are done with an 'RT' tag, representing retweet. Also, an '@' symbol followed by a user's name is used to mention a particular user in a tweet. Likewise, the hashtag symbol, '#' can be used to make mentions or comments. An example of a re-tweet that makes mention of another user is: "RT @user: I also like the new treatment plan initiated by #Mayo Clinic".

The content of discussions dedicated to MDD on Twitter and how it relates to the influence of users is the basis of analysis in case study demonstrated in this paper.

Major Depressive Disorder

MDD is a leading cause of disability in the U.S. (Lopez and Murray 1998) and requires prolonged management or life-time treatment with associated side effects. According to the Center for Disease Control (CDC), at least 9.5 percent of the U.S. adult population suffer from depression each year (Center for Disease Control and Prevention 2013). Individuals who suffer from depression are more likely to experience reduced productivity due to missed workdays. CDC reports that in a 3-month period, depression patients suffer 11.5 days of reduced productivity as a result of missed workdays (Center for Disease Control and Prevention 2013). The associated cost to employers for instance, as a result of missed workdays range between $17 to $44 billion dollars (Center for Disease Control and Prevention 2013).

The prolonged nature of the disease creates a need for active participation by patients themselves in managing it. New approaches for self-managing MDD are therefore needed and should include patients' perspectives and collective wisdom generated from the Internet. The proliferation of electronic SNs offer an effective and objective platform for investigating the perspectives of patients on how to manage this and other chronic diseases.

CRISP-eSNeP Process Development

In this section, we introduce our proposed process (CRISP-eSNeP) for managing and analyzing large data on electronic SNs (figure 2) as an extension to the CRISP-DM process. We present our arguments through an analysis of Twitter data on MDD, where we describe our experiences and lessons learned.

With reference to Hevner et al. (2004) information systems research framework which combines design science and behavioral science approaches, we develop a set of guidelines for managing electronic SN data (table 5) as part of our process. Our guidelines, termed as Cross Industry Standard Process for Electronic Social Network Platforms (CRISP-eSNeP), is an extension to the CRISP-DM methodological process. The process comprises multiple phases, the core of which are Data Acquisition, Data Cleaning, Data Formatting, Data Validation, Data Analysis and Result and Deployment. The process also offers an iterative transition in-between phases. In addition to the six core phases, some implementations may require an initial need to develop a big data management platform that is unique to an environment such as the Twitter platform. We describe our process in the subsequent sections.

Data Acquisition

This phase involves accessing the needed data in a timely manner. Even though social network platforms are avenues to access large volumes of data, most big data implementations struggle to acquire the right data in a timely manner. In a recent study, only 36% of big data implementations had access to timely data that was capable of supporting the generation of new insights and knowledge (Tanner Jr. 2014). Also, some research questions warrant the combination of data from multiple sources. Acquisition of data from SNs may require the use of high capacity data platforms to extract and store data in a timely manner.

There are various methods for extracting relevant data from Twitter and other SN websites. A list of keys words relevant to topic of interest can be used to extract the data via the Twitter API. An example of the use of such a method was presented by Choudhury et al. (2013) in a study in which via crowd sourcing, health-related information from a SN platform was collected. Data acquisition can be done using Apache Flume, a sub-project of the Apaches Hadoop platform. Most major big data analytics tool providers include facilities for extracting data from social networks.

Specific vendor tools employ the API provided by the social network to extract the data. In our case, this data was collected using Apache Flume, a subproject to Apache Hadoop platform. The tweets were marked for initial collection in these contained MDD related keywords. MDD-related keywords used were based on the open source UMLS medical database, (UMLS Reference Manual 2009) and from clinical experts. To identify an MDD-related tweet, the tweet must contain an MDD-related keyword. Also, words which may be MDD-related, yet rather used in common language in ways that does not relate to MDD were identified and removed. For instance, a tweet such as "school feels depressing today", even though contains the word "depressing", may have no relation to an MDD symptom. The keywords defined for data extraction were categorized into three main groups; medications (trade and generic names), side

effects, signs and symptoms. Using the Hadoop platform, we collected tweets for 24 hours per day over a four week period. The original volume of data was about 170 gigabytes.

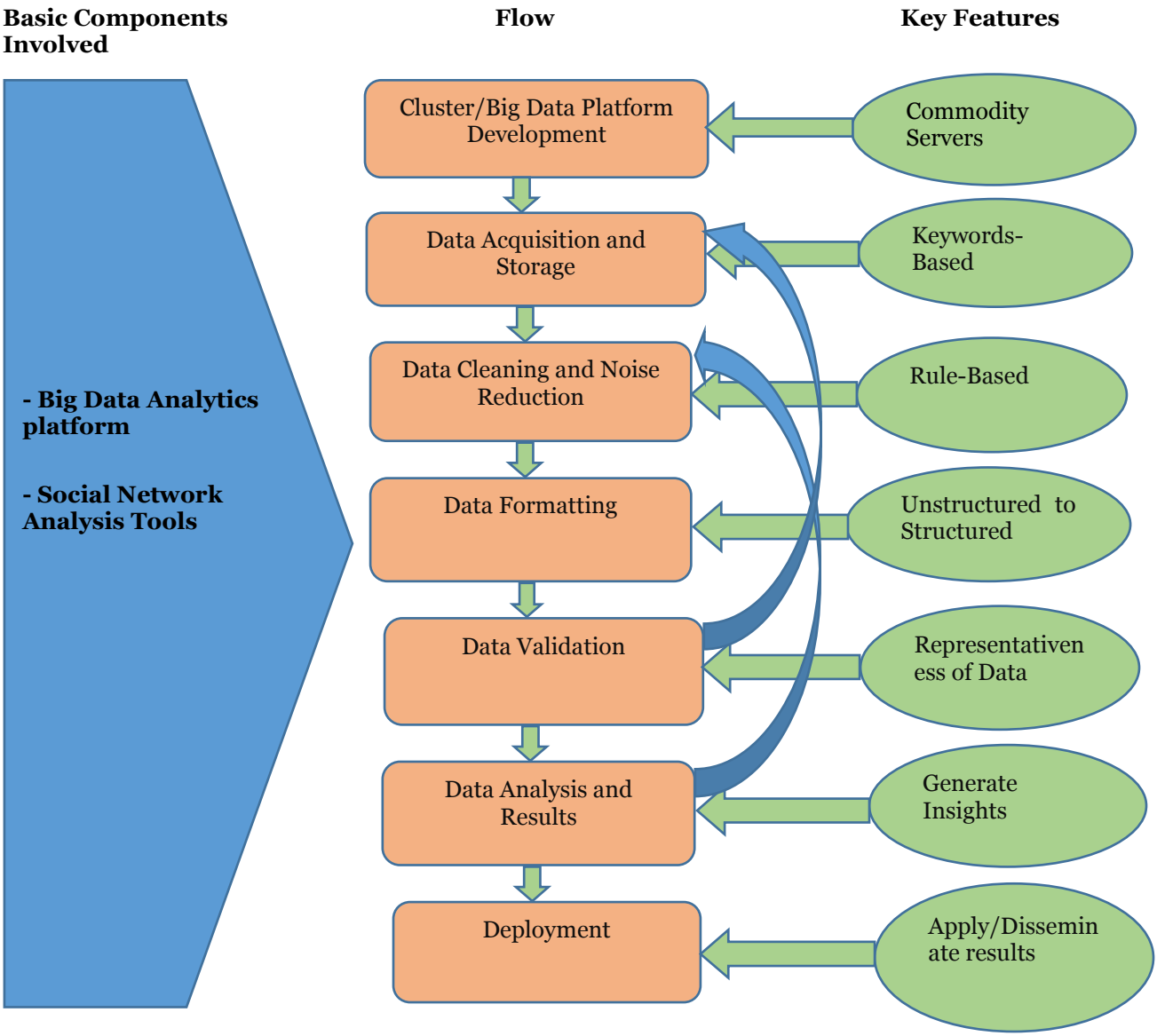

Figure 2. CRISP-eSNeP Model for Data Management and Analysis

Data Cleaning

The data cleaning phase is the stage where non-relevant data regarding the research question is eliminated. Several strategies for handling 'dirty" data including error detection methods and data repairing algorithms have been developed in recent studies (Tang 2014). More is only better if the data acquired is devoid of unnecessary data elements that would skew the results of the analysis. In large data sets, automated processes based on rule sets may be utilized to eliminate unwanted data.

In our study, we filtered the extracted data based on a combination of keywords. That is, if a tweet contained a certain combination of keywords, it was retained in the data set. This helped select more relevant tweets that had MDD-related words. The category of keywords created were medications (both trade and generic names) side effects, signs and symptoms and alternate treatments (snapshot in table 1).

Primary List (Trade and Generic names) - L_A	**Side Effects – L_B**	**Signs and Symptoms – L_C**	**Alternatives – L_D**
Antidepressants, SSRI, TCA, Pristiq, Cymbalta, Fetzima	diarrhea, nausea, restlessness	fatigue, moody, depression	Yoga, meditation, God

Table 1. Categories of Keywords Used for Data Extraction

A set of four rules were applied after data extraction to further clean the data (table 2). A word could belong to list L_A, L_B, L_C, or L_D. For instance for rule 1, a tweet that contains a word in list L_A was considered viable. Also, a tweet that contains words in both L_B and L_C was viable. Tweets that contain words only in L_B or L_C were not considered viable. For example, if a tweet contains just the word 'helpless' (in L_C), it is not considered viable since it is considered too broad a description for MDD. However, if the same tweet also contain the word 'insomnia', it is assumed that the tweet is likely an MDD discussion.

Rule	**Expression**
Rule 1	L_A Only
Rule 2	$L_B \cap L_C$
Rule 3	$L_B \cap L_D$
Rule 4	$L_C \cap L_D$

Table 2. Data Filtration Rule Set

Data Formatting

Current sources of big data, especially via SNs generate unstructured data that cannot be readily analyzed by existing data analytics platforms. These data sets could come in various formats including text and video. This phase requires that unstructured data sets are converted into structured formats prior to analysis. It requires that raw data is curated to the extent that relevant staff, even if non-technical can perform necessary analytical functions (Terrizzano et al. 2015).

The raw data generated from Twitter is classified as unstructured since it does not have a pre-defined data model. In order to improve the speed of data extraction and storage on the Apache Hadoop platform, we first stored the data in an unstructured format. Subsequently, a schema was applied to it prior to retrieval and query. Hence, whereas data storage in the HDFS is fast and optimized for large data storage, data retrieval and query is relatively slow. The formatted data, stored on HDFS was queried using Apache Hive, an SQL-like language.

Data Validation

In this phase, the accuracy of the data is ascertained based on the source of data and related research question. Access to a huge amount of data does not necessarily mean access to knowledge. The large amount of data needs to be representative of the entire population being studied and should exhibit the characteristics of the problem domain. For instance, on SNs, the data acquired data should exhibit a power-law characteristic (Lerman and Ghosh 2010) if the research question requires analysis of the influence wielded by networks nodes.

In this phase, we validated the social graph by assessing the core strength of the Twitter network we extracted. Mislove et al. (2007) describe 'core of a network' as satisfied by two main conditions;

1) Connectivity of all parts of the network

2) Relatively small diameter of core as compared to the diameter of the whole graph.

The rationale behind these conditions is that since a minor portion of the network controls the information flow on the network, removal of these core set of users should lead to a more disperse network in which users are more distant from each other. The highly connected users at the core of the network serve as connection points between other less connect users. The network core contributes to shorter path lengths and fewer edges between nodes (Mislove et al. 2007).

Graph Compactness	Node Characteristic			
		Percentage of Nodes Considered (%)	**Percentage of Edges Considered (%)**	**Percentage of Shortest Paths Considered (%)**
	Full Graph	100	100	100
	Reduced Graph1	99.97	62.69	63.28
	Reduced Graph2	99	52.72	53.4

Table 3. Levels of Graph Compactness

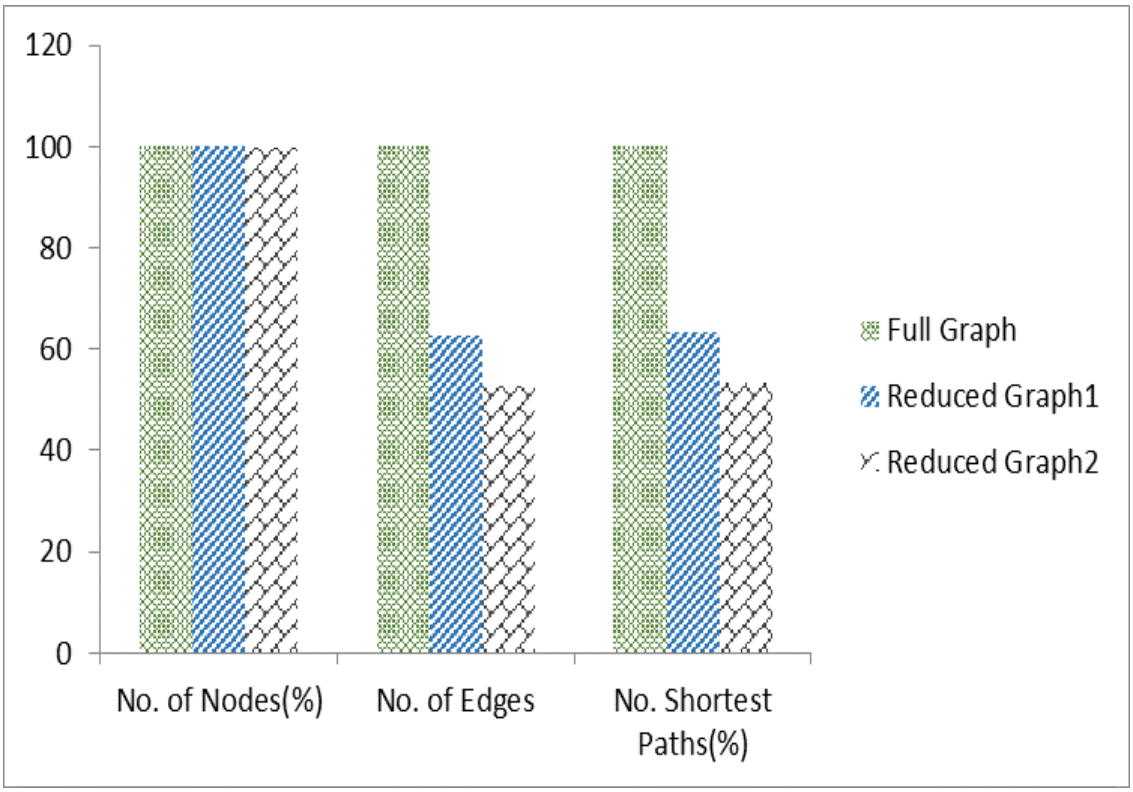

Figure 3. Percentages of Number of Edges and Shortest Paths

A path length is described as the distance between any two nodes. Central users serve as transition points between other nodes and offer shorter distances between other nodes. When central users are eliminated, nodes go through longer routes in order to connect to their destination. Based on a total graph size of 51, 424 nodes, the number of highly connected nodes that formed the core of the network were decreased in

two steps. First, we removed less than 1% of the nodes (belonging to core nodes) in the total graph. This contributed to about 37% decrease in the number of edges and shortest paths between nodes (shown in table 3). This implies that as the number of nodes that held the graph together were decreased, interconnection between users were widened. Next, we decreased the number of nodes (belonging to core nodes) in the graph by one percent. The result showed a further 10% decrease in the number of edges and shortest paths between the remainder of the nodes (figure 3). This analysis indicates that the extracted graph is representative of a SN since it has a densely connected core, without which the SN would disintegrate.

Data Analysis and Results

During this phase, the acquired big data is analyzed iteratively based on research questions posed. For agile and deep analysis, this phase calls for advanced data-parallel statistical algorithms using platforms such as Hadoop (Cohen et al. 2009). The results and implications of the study are also presented at this phase.

We operationalized both influence and quality. We recognize that use of a large sample size could potentially increase the erroneous occurrence of statistical significance (Lin et al. 2013), hence leading to inaccurate conclusions. Previous research have suggested ways to address this problem. For instance, (Dubitzky et al. 2007) suggested the use of random permutation tests when dealing with large sets of consisting of numerous variables. In our study, we accounted for this potential problem in two ways. First, we chose groups of conversations as our unit of analysis rather than individual tweets. We created a total of hundred groups of tweet conversations rather than exposing the entire individual tweets in our corpus to significant testing. Also, once we collected and cleaned the initial data set, we ensured that only a subset that was representative of the social graph was used for significance testing.

Analyzing Influence

The analysis was informed by our project research question, which asked whether the influence of a node was related to the quality of messages disseminated by that node. On Twitter, as a node gets more reference from other adjacent nodes, it implies that the node is important and can attract traffic and attention more than it gives. This phenomenon makes the node influential in the network. To measure influence of a node 'a', we took into account the sum of nodes that are directed from adjacent nodes to node 'a'. We then deducted from it the sum of nodes that node 'a' is directed to as shown in equation. The difference indicates the influence of node 'a' in the network as is shown in equation 1.

$$I_a = \sum_{i=1}^{n} Gi \; - \sum_{i=1}^{n} Yi = ID_a - OD_a \qquad (1)$$

Where:

I_a = *Influence of node 'a'* G_i = *Incoming link to 'a'*

Y_i = *Outgoing link from 'a'* ID_a = *Sum of node 'a' in-links*

OD_a = *Sum of node 'a' out-links*

We performed this analysis for two set of nodes; central nodes and non-central nodes. In this study, we define a central node as a node that has more than 100 friends or adjacent connections and a non-central node as a node that has less than 100 friends. The focus of the influence analysis was on the number of friends for each node which is a better reflection of a node's influence than number of followers (Weiss 2009). This number is generally reported as about 50% of the number of followers per node. Given that the average number of followers per Twitter user is about 208 (Smith, 2014), we chose a conservative value of 100 as the break point for the centrality of a node on Twitter (Roberts 2012).

Analyzing Quality

We determined quality (Q_t) at the group level based on a tweet's lexical structure using Vosecky, Leung, & Ng's (2012) 'formality score' method shown in equation 2. Quality measures, calculated as a lexical score, were based on statistical counts of different parts-of-speeches present in the tweet. For instances, in a properly constructed sentence, there are more descriptive words (adjectives) and nouns which indicate

the diversity of the lexical structure of the sentence. In a conversation t, consisting of multiple tweets, T, the conversation, t can be expressed as:

$$t = \{T_1, T_2, \ldots T_N\} \quad (2)$$

A tweet, T will have one or more elements in terms of words, W and can be expressed as:

$$T = \{W_1, W_2, \ldots, W_N\} \quad (3)$$

Again, we tokenized the tweets belonging to a set of tweet conversation. By means of parts-of-speech (PoS) tagging, we classified all words into different parts-of-speeches. A frequency count of each of the tagging properties were generated. For example, the frequency of pronouns in a conversation, t is given as:

$$FP_t = \sum_{i=0}^{n} W_P \quad (4)$$

Where W_P represents counts of an instances where a particular word is tagged as a pronoun.

Finally, quality based on the lexical score (Q_t) used is depicted by equation 5;

$$Q_t = \sum_{i=0}^{n} ((FN_t + FAdj_t + FPre_t + FNG_t - FV_t - FAdv_t - FIntj_t + \lambda)/2) \quad (5)$$

Where:

Q_t = Quality of information

FN_t = Noun

$FAdj_t$ = Adjective

$FPre_t$ = Preposition

FNG_t = Noun Group

FV_t = Verb

$FAdv_t$ = Adverb

$FIntj_t$ = Interjection

In the original '*formality score*' method (Lahiri et al. 2011) where longer sentences were analyzed, λ was set as 100. However, for a shorter standard tweet of 140 (maximum) we set λ as 10 (Vosecky et al. 2012). Based on the formula in equation 5, we performed a t-test using the Welch-Satterthwaite method to test the difference in quality between influential and non-influential groups of users.

Results of Analysis

We determined the formality score (Q_t) for both influential and non-influential users groups. Table 4 shows the results of the Welch-Satterthwaite method to test significant difference in quality of messages initiated by influential and non-influential nodes. The table indicates that quality of information disseminated by influential users was significantly higher than that disseminated by non-influential users ($t(55) = 2.004$, $p < 0.05$).

Group	Number of Discussions	Mean Quality	Standard Deviation	Standard Error	P Value
Influential Users	50	74.17	20.54	2.90	P = .00
Non-Influential Users	50	11.08	4.98	0.705	

Table 4. Effect of Influence on Quality

Furthermore, we performed post-hoc tests to verify the results we had during the analysis phase. Using a regression test, we checked if influence was a significant predictor of information quality. We also performed text analysis of the actual tweets posted to identify deeper trends of conversations among both influential and non-influential users. We used the same set of keywords used in the data cleaning phase to categorize tweets among both influential and non-influential nodes. This was done to understand the trend of information that were posted on Twitter.

Deployment

As is proposed by (Chen et al. 2012a), identification of interesting results is of little use unless the results can be operationalized and implemented within one or multiple domain areas. The deployment could also involve outputs such as the generation of a report or the implementation of a business strategy to help improve a business process.

Specifically, the results from this case study will help in the design of strategic initiatives on SNs that will support disease self-management efforts by both health practitioners and patients.

Implications

There are multiple implications of this methodology to both theoretical research and analytics practice. First, the methodology offers a concise process for theoretical research as it encourages the development of conceptual and analytical questions. These questions then drive the course of the analysis while identifying valuable insights from the data. The methodology could also help streamline the arduous tasks of managing high volume unstructured data that has the potential of providing meaningful insights, yet could crumble when there is no concise data analytics process to manage it. A set of analytics guidelines for SNs based on our CRISP-eSNeP method is listed in table 5.

CRISP-eSNeP Phase	Guideline
Data Acquisition	An initial keywords list based on domain expertise should be developed to extract data at a manageable level using big data tools such as Hadoop.
Data Cleaning	Further cleaning of SN data can be based on compound or nested filtering to retain relevant content.
Data Formatting	Convert unstructured data into a structured format set for further analysis using tools such as Apache Hive.
Data Validation	Verify that the dataset exhibits characteristics of the population under study. In SNs, this could entail checking for a power-law distribution.
Data Analysis	Analysis using relevant analytics and statistical methods should be based on a set of previously identified questions. Interpretation of significance testing should take into account the size of data set used.
Deployment	Well-documented and actionable results should be applied to problem of interest.

Table 5. Guidelines for CRISP-eSNeP Based on CRISP-DM

Conclusion

Even though the domain in which we presented our argument is in healthcare, this methodological process can be applied to other domains where SN analytics is employed. Both current and historical data generated by businesses on their business operations have becomes increasingly challenging to analyze as a result of their sheer size. On electronic SNs, our CRISP-eSNeP methodological process can be applied to manage, analyze and generate valuable insights.

When data is acquired from SN platforms, it must be ensured that the mechanism of message propagation is considered. For instance, the Twitter platform operationalizes a directed network whereas Facebook allows for a non-directed network. Direction and mechanism for data propagation should, therefore, capture such peculiarities. The Twitter social graph utilized in this study is only a model network. Lessons learned can illuminate the understanding of health-related big data analytics on other SNs.

Extensions of this research will test the use of this methodology in other domain areas such as retail. We will also study how this data analytics process can support research using other research principles besides social science.

REFERENCES

Center for Disease Control and Prevention. 2013. "Workplace Health Promotion," (available at http://www.cdc.gov/workplacehealthpromotion/implementation/topics/depression.html).

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quaterly* (36), pp. 1165–1188.

Choudhury, M. De, Gamon, M., Counts, S., and Horvitz, E. 2013. "Predicting Depression via Social Media," *Association for the Advancement of Artificial Intelligence* (2) (available at http://research.microsoft.com/EN-US/UM/PEOPLE/horvitz/depression_social_media_icwsm_2013.pdf).

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., and Welton, C. 2009. "MAD Skills: New Analysis Practices for Big Data," *Proceedings of the VLDB Endowmen* (2), pp. 1481–1492 (doi: 10.14778/1687553.1687576).

Dean, J., and Ghemawat, S. 2008. "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM* (51), pp. 1–13 (doi: 10.1145/1327452.1327492).

Dubitzky, W., Granzow, M., and Berrar, D. 2007. *Fundamentals of Data Mining in Genomics and Proteomics*, Springer Science & Business Media (doi: 10.1007/978-0-387-47509-7).

Gandomi, A., and Haider, M. 2015. "Beyond the Hype: Big Data Concepts, Methods, and Analytics," *International Journal of Information Management* (35:2), pp. 137–144.

Heidemann, J., Klier, M., and Probst, F. 2012. "Online Social Networks: A Survey of a Global Phenomenon," *Computer Networks*, pp. 3866–3878 (doi: 10.1016/j.comnet.2012.08.009).

Hesse, B. W., O'Connell, M., Augustson, E. M., Chou, W.-Y. S., Shaikh, A. R., and Rutten, L. J. F. 2011. "Realizing the Promise of Web 2.0: Engaging Community Intelligence.," *Journal of health communication. International Perspectives* (16:1), pp. 10–31 (doi: 10.1080/10810730.2011.589882).

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information System Research," *MIS Quarterly* (28:1), pp. 75–105.

Hu, W., Xie, N., Li, L., Zeng, X., and Maybank, S. J. 2011. "A Survey on Visual Content-Based Video Indexing and Retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* (41), pp. 797–819 (available at http://eprints.bbk.ac.uk/5568/).

Hughes, B., Joshi, I., and Wareham, J. 2008. "Health 2.0 and Medicine 2.0: Tensions and Controversies in the Field.," *Journal of Medical Internet Research* (10:3) (doi: 10.2196/jmir.1056).

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. "What is Twitter , a Social Network or a News Media?," in *Proceedings of the 19th International Conference on World Wide Web. ACM*, Raleigh, North Carolina.

Lahiri, S., Mitra, P., and Lu, X. 2011. "Informality Judgment at Sentence Level and Experiments with Formality Score," in *In Computational Linguistics and Intelligent Text Processing*, Berlin Heidelberg: Springer, pp. 446–457.

Lerman, K., and Ghosh, R. 2010. "Information Contagion: A n Empirical Study of the Spread of News on Digg and Twitter Social Networks," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 90–97.

Lin, M., Lucas, H. C., and Shmueli, G. 2013. "Research Commentary —Too Big to Fail: Large Samples and the p -Value Problem," *Information Systems Research* (24), pp. 906–917 (doi: 10.1287/isre.2013.0480).

Lopez, A. D., and Murray, C. C. . J. . 1998. "The Global Burden of Disease , 1990 – 2020," *Nature Medicine* (4), pp. 1241–1243 (doi: 10.1038/3218).

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. 2007. "Measurement and Analysis of Online Social Networks," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement IMC 07* (Vol. 40), San Diego, U.S.A, pp. 29–42 (doi: 10.1145/1298306.1298311).

Roberts, J. J. 2012. "Typical Twitter User is a Young Woman with an iPhone & 208 followers," (available at https://gigaom.com/2012/10/10/the-typical-twitter-user-is-a-young-woman-with-an-iphone-and-208-followers/).
Sagiroglu, S., and Sinanc, D. 2013. "Big Data: A Review," in *International Conference on Collaboration Technologies and Systems (CTS)*, pp. 42–47 (doi: 10.1109/CTS.2013.6567202).
Shearer, C. 2000. "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing* (5), pp. 13–22.
Shi, Z., Rui, H., and Whinston, A. 2014. "Content Sharing in a Social Broadcasting Environment: Evidence from Twitter," *MIS Quarterly* (38), pp. 123–142 (doi: 10.2139/ssrn.2341243).
Singh, J. 2014. "Apache Hadoop Ecosystem," (available at http://techblog.baghel.com/index.php?itemid=132).
Tang, N. 2014. "Big Data Cleaning," in *Web Technologies and Applications*, Springer International Publishing, pp. 13–24.
Tanner Jr., J. F. 2014. "Big Data Acquisition," in *Analytics and Dynamic Customer Strategy*, Hoboken, NJ: John Wiley & Sons, Inc, pp. 85–101 (doi: 10.1002/9781118919767.ch5).
Terrizzano, I., Schwarz, P., Roth, M., and Colino, J. E. 2015. "Data Wrangling: The Challenging Journey from the Wild to the Lake," in *7th Biennial Conference on Innovative Data Systems Research (CIDR '15)*, Asilomar, Carlifornia.
The Apache Software Foundation, F. 2014. "Welcome to Apache™ Hadoop," (available at http://hadoop.apache.org/).
Tinati, R., Halford, S., Carr, L., and Pope, C. 2014. "Big Data: Methodological Challenges and Approaches for Sociological Analysis," *Sociology* (48), pp. 663–681 (doi: 10.1177/0038038513511561).
UMLS. 2009. "UMLS® Reference Manual [Internet]," (available at http://www.ncbi.nlm.nih.gov/books/NBK9676/).
Vosecky, J., Leung, K. W.-T., and Ng, W. 2012. "Searching for Quality Microblog Posts: Filtering and Ranking Based on Content Analysis and Implicit Links," in *Database Systems for Advanced Applications*, Springer Berlin Heidelberg, pp. 397–413 (available at http://link.springer.com/chapter/10.1007/978-3-642-29038-1_29).
Weiss, T. 2009. "Twitter – Do Friends Count More than Followers?," (available at http://www.trendsspotting.com/blog/?p=608).
Wirth, R. 2000. "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39 (doi: 10.1.1.198.5133).
Wu, X., Zhu, X., Wu, G.-Q., and Ding, W. 2014. "Data Mining with Big Data," *IEEE Transactions on Knowledge and Data Engineering* (26:1), pp. 97–107 (doi: 10.1109/TKDE.2013.109).