Original Paper

Exploring Big Data Analytic Approaches to Cancer Blog Text Analysis

Viju Raghupathi¹, PhD; Yilu Zhou², PhD; Wullianallur Raghupathi³, PhD

¹Associate Professor, Brooklyn College of the City University of New York, NY, USA
²Professor, Gabelli School of Business, Fordham University, NY, USA
³Associate Professor, Gabelli School of Business, Fordham University, NY, USA

Corresponding author:
Wullianallur Raghupathi, PhD.
Professor, Gabelli School of Business
Fordham University
140 W. 62nd Street
New York, NY 10023
USA
Email: raghupathi@fordham.edu
Exploring Big Data Analytic Approaches to Cancer Blog Text Analysis

Abstract

Background: In recent years researchers have begun to realize the value of social media as a source for data that helps us understand health-related phenomena. Health blogs in particular are rich with information for decision-making. While there are web crawlers and blog analysis software that generate statistics related to blogs, these are relatively primitive and are not useful computationally to aid with the analysis and understanding of the social networks and medical blogs that are evolving around healthcare. There is a need for sophisticated tools to fill this gap. Furthermore, to our knowledge there are not many big data studies or applications in the text analytics of cancer blogs. This study attempts to fill this specific gap while analyzing cancer blogs.

Objective: In this exploratory research, we examine the potential of applying big data analytic techniques to the analysis of blogs that exist in the cancer domain. Our objective is twofold: to extract from the blogs, patterns and insight about cancer diagnosis, treatment, and management; and to apply advanced computation techniques in processing large amounts of unstructured health data.

Methods: We applied the big data analytics architecture of Hadoop MapReduce via the Cloudera platform to the analysis of cancer blog content, in order to extract patterns and insight on cancer diagnoses. We apply a series of algorithms to gain insight into the content and develop a vocabulary and taxonomy of keywords based on existing medical nomenclature. By applying a number of algorithms, we gained insight into the blog content. The study identifies, for instance, the most discussed topics as well as associations that relate to key phenomena.

Results: Using several text analytic algorithms, including word count, word association, clustering, and classification, we were able to identify and analyze the patterns and keywords in cancer blog postings. This gave insight into some of the key issues that are discussed in blogs such as the type of cancer (breast cancer being the dominant topic), diagnosis, treatments, and others.

Conclusions: In general, big data analytics has the potential to transform the way practitioners and researchers gain insight from health social media, especially those in free text, unstructured form. Big data analytics and applications in health-related social media are still at an early stage, and rapid acceleration is possible with the advancements in models, tools, and technologies.

Keywords: big data analytics, text analytics, cancer blog, Hadoop, MapReduce

In recent years researchers have begun to realize the value of social media as a source for data that helps us understand health-related phenomena1, 2. Numerous past and ongoing studies as well as applications, have applied a range of techniques (including statistical, machine learning, and visualization) to structured and
unstructured social media health data to perform sentiment analysis, elicit patterns, and provide decision support. The social media content includes that found in tweets, blogs, and web search logs, among others. The healthcare domain has seen a tremendous increase in the use of Web 2.0 tools and social media such as blogs, wikis, podcasts, twitter feeds, vlogs (video blogs) and on-line journals that convey health-related information. These and other content-driven applications enable physicians, patients, hospitals, insurance companies, government, and others—key participants in the health care system—to create and disseminate health information via the web. Patients, for example, need only put health-related terms into Google Search to find useful information related to diagnosis, treatment, and the management of diseases. This development suggests the enormous potential of online media to inform and improve personalized medicine and population health management. Physicians, too, use such tools to conduct research in the context of evidence-based medicine and to address patients’ concerns and issues. Hospitals and other providers use these tools as “gateways” to the communities. As large repositories of unstructured textual data emerge and grow, health entities are examining the potential of text analytics and other methods to evaluate the data and glean patterns and relationships. These patterns and relationships are, in turn, assessed to gain insights for making informed health decisions and improving clinical outcomes. Text mining bridges the gap between free-text and structured representation of cancer information. Text mining uses techniques from natural language processing (NLP), knowledge management, data mining, and machine learning (ML) to process large document collections. These techniques support information retrieval, (which gathers and filters relevant documents), as well as document classification, (which maps documents to appropriate categories based on their content), information extraction (which selects specific facts about pre-specified types of entities and relationships of interest), terminology extraction (which collects domain relevant terms from a corpus of domain-specific documents), named entity recognition (which identifies entities from predefined categories), etc.

Health data, such as general patient profiles, clinical data, insurance data, and other medical data, are being created for various purposes, including regulatory compliance, public health policy analysis and research, and diagnosis and treatment. Data may include both structured data (e.g. patient histories as records in a database) and unstructured data (e.g. audio/video clips, textual information such as in blogs or physician’s notes). Text analytics is typically used to identify patterns and trends in the unstructured data. These patterns can shed light on a wide range of issues such as drug reactions, side effects, treatment outcomes, personalized medical treatments, and efficacy of drugs. One famous example of analytics shedding light on a medical mystery was the discovery of an association between the arthritis drug Vioxx and an increased risk of heart attack/stroke, resulting in the withdrawal of the drug from the market.

With regard to health social media analytics, several papers have examined the potential. A total of 285,417 Twitter posts, also known as tweets, were analyzed about HPV vaccines. These were from 101,519 users, whose total followers numbered some 4,387,524 individual accounts. The goal of the study was to evaluate
the use of community structure and topic modeling methods (methods for discovering the abstract concepts that occur in a corpus of documents), as a process for characterizing the clustering of opinions about human papillomavirus (HPV) vaccines on Twitter. The authors tested Latent Dirichlet Allocation and Dirichlet Multinomial Mixture (DMM) models for inferring topics associated with tweets. This was followed by the application of community agglomeration (Louvain) and the encoding of random walk (Infomap) methods “to detect community structure of the users from their social connections”\(^5\). They examined the alignment between community structure and topics using several common clustering alignment measures, and they introduced a statistical measure of alignment based on the concentration of specific topics within a smaller number of communities. They concluded that the use of community detection in concert with topic modeling appears to be a useful way to characterize Twitter communities for the purpose of opinion surveillance in public health applications. Their approach may help identify online communities at risk of being influenced by negative opinions about public health interventions such as HPV vaccines.

Research has focused on identifying the quality of hospital service automatically using online communities\(^24\). The authors defined social-media-based quality factors for hospitals. In addition, they developed text-mining techniques to detect such factors as professionalism, process, environment, and impression that frequently occur in online health communities. Then, after identifying factors that represent qualitative aspects of hospitals, they applied a sentiment analyses to recognize types of recommendations in messages posted within online health communities. Others examined the potential for post marketing safety surveillance from patient experiences with drugs reported in social media\(^25\); and discussed how user content posted on Twitter is subject to bio surveillance: to characterize public perception of health-related topics and as a means of distributing information to the general public\(^26\). A study examined large amounts of unstructured, free-text information on blogs, social networks, and physician ratings websites that describe the quality of healthcare available\(^2\). The authors used sentiment analysis techniques to categorize online free-text comments by patients as either positive or negative descriptions of the healthcare they received\(^2\). From these free-text descriptions, the authors attempted to automate predictions as to whether a patient would recommend a hospital, whether the hospital was clean, and whether they were treated with dignity and to compare those automated predictions to the patient’s own quantitative rating of their case\(^2\). The goal was to improve the overall quality of healthcare.

Studies have assessed and explained the online use of alcohol-related Chinese keywords and validated blog searching as an infoveillance method for surveying changes in drinking patterns in Hong Kong\(^6\). Others have examined the role of social media in bio surveillance application\(^1\). For instance, the authors mined for influenza mentions as well as for information dissemination and public sentiment towards such topics as vaccination.

Health forums provide rich, detailed content of the patient experience vis-à-vis various health issues, including temporal and emotional factors that may help us tailor information to fit their needs. Such extracted information can also be used for
decision support and to enrich bio surveillance and biomedical research\textsuperscript{27}. Studies have developed algorithms that alert to possible outbreaks of communicable diseases using Internet data - specifically Twitter and search engine queries\textsuperscript{7}. Other studies have conducted surveillance and analysis of Twitter data to characterize the frequency of non-medical use of prescription medications (NUPM) and identify illegal access to drugs via online pharmacies\textsuperscript{3}. Researchers show how the logs of queries submitted to search engines could potentially be sources for the detection of emerging influenza epidemics particularly when changes in the volume of search queries are detected (infodemiology); they then describe a methodology for detecting influenza outbreaks using search query data\textsuperscript{28}.

There are extensive discussions on how recent studies have demonstrated the utility of social media data sources to address a wide range of public health goals\textsuperscript{29}. These include epidemiological surveillance systems for influenza, allergies, tracking health behaviors (such as smoking and exercise), identifying mental health trends, as well as measuring health perceptions and sentiment. In yet another study, results from the authors’ experiments suggest that it is possible to accurately predict future patient visits from geotagged mobile search logs\textsuperscript{8}. The web queries of millions of anonymized searchers have been statistically analyzed\textsuperscript{30}. The specific goal was to mine signals from large-scale Web search logs of symptom queries for pancreatic adenocarcinoma. Their results highlight the promise of using Web search logs as an innovative direction for screening for pancreatic carcinoma.

Data collected from online search logs was analyzed to better understand the relationship between patterns in the seeking of online health information and healthcare utilization (HU)\textsuperscript{31}. For instance, users’ online search and access activities were looked at, before and after queries that sought medical professionals and facilities. The results provided insight into how users decide whether and when to utilize healthcare resources and what effect health concerns and professional advice have on seeking healthcare facilities and service. The research extracted queries originating in the U.S. via the Bing search engine. Batches of pharmaceuticals are sometimes recalled from the market when a safety issue or a defect is detected in specific production runs of a drug. The study tested the hypothesis that defective production lots can be detected earlier by monitoring queries to Internet search engines. Findings suggested aggregated Internet search engine data can be used to facilitate early warnings of faulty medicines\textsuperscript{32}.

Health blogs in particular are rich with information for decision-making. While there are web crawlers and blog analysis software that generate statistics related to blogs (including, say, the number of blogs or Top Ten blogs in a certain category), these tools are relatively primitive and are not useful computationally to aid with the analysis and understanding of the social networks and medical blogs that are evolving around healthcare. Thus, there is a critical need for sophisticated tools to fill this gap. Furthermore, to our knowledge there are not many big data studies or applications in the text analytics of cancer blogs. This study attempts to fill this specific gap while analyzing cancer blogs.

Prior research has applied traditional machine learning techniques to this type of analysis. But considering the volume of data and the fact that it is growing exponentially and in a variety of data types, new and dynamic approaches are
needed to analyze these substantial amounts of unstructured text data\textsuperscript{33}. In this exploratory research, we examine the potential of applying big data analytic techniques, including Hadoop MapReduce, to the analysis of blogs that exist in the cancer domain. Our objective is twofold: to extract from the blogs, patterns and insight about cancer diagnosis, treatment, and management; and to apply advanced computation techniques in processing large amounts of unstructured health data. The big data platform Hadoop MapReduce has the potential to address the limitations of prior approaches by distributing the storage and processing of large amounts of data. Scalability, richness of algorithmic application, speed, and robustness are among the potential benefits\textsuperscript{33,34}.

The rest of this article is organized as follows: in the following section we provide background for the research by describing the domain of cancer blogs and by explaining the big data analytics Hadoop MapReduce and its methodology; next, we describe our research objectives and methodology in the text analytics of cancer blogs; we then discuss our results; this is followed by scope and limitations of our study; and finally, we offer conclusions and suggestions for future research.

Research Background

Blogs

Blogs (short for web logs) are web pages that often resemble personal diaries and contain entries or posts, typically in reverse chronological sequence\textsuperscript{35}. (In other words, an author’s most recent blog entry is the first one you read when you visit the site.) Blogs are powerful vehicles for sharing information and voicing opinions on a vast range of issues and causes, including but not at all limited to those that lean toward economics, politics, medicine, society, and the personal\textsuperscript{36}, and cover the detail spectrum from complex and technical to trivial. Blogs see various levels of user participation. Some users participate passively by only reading content, while others are more active, posting comments or contributing text\textsuperscript{36} when blog administrators permit it.

As described earlier, analyzing health blogs can lead to insights related to diseases and treatments (e.g. alternative medicine, therapy), as well as provide support links. For example, such analysis can reveal the most common issues patients have, the types of diseases that are most commonly discussed and why, the kinds of therapies and treatments discussed most often, and any medical or non-medical information provided. In addition, the information can reveal profiles of and patterns relating to the bloggers themselves - identifying, for instance, which bloggers offer relevant and accurate information and which major factors motivate the postings.

Analyzing the content in blogs (text analytics) requires efficient methods for indexing, codifying, and managing data. However, the analysis and interpretation of health-related blogs is challenging for a number of reasons. First, while blogs enable the formation of social networks of patients and providers, the prolificacy of the health/medical terminology comingingled with the subjective vocabulary of the patient can be a challenge to interpret. Second, the blog world is characterized by an absence of rules on the format, method of posting, and structure of the content,
which leads to variability in word choice, sentence structure, grammar, and punctuation. Meanwhile, embedded in this free-form text are bits and pieces of important information, which, if aggregated and summarized, could correlate to intelligent responses to integral medical questions. Third, the content in health blogs incorporates two important psychological facets of the bloggers: their feelings and their mindset in terms of how they are managing their cancer. These meaningful aspects necessitate an element of sentiment analysis. Fourth, the absence of cues that would be present in face-to-face conversations compounds the challenge of interpretation. In a face-to-face exchange, two parties not only send and receive information, but they also interpret information based on body language, facial expressions, vocal cues, and an understanding of each other’s circumstances. Fifth, a comment from a blogger cannot be viewed in isolation from its original post and preceding comment. Analyzing a post without considering the thread it relates to means ignoring contextual clues undermining the meaning of the post. Sixth, the amount of available blog content is rapidly expanding, necessitating the creation of more complex and sophisticated techniques for data preparation and analysis.

In the face of such challenges, researchers need a strong framework to analyze the content of health blogs thereby assisting in clinical decision-making and reducing the cost of overall healthcare delivery. Two high-level queries can arise from text mining of healthcare blogs.

**How can we make sense of the aggregate healthcare content?**

**And how can one perform a meta-analysis to interpret and generalize healthcare content in terms of patterns of diagnosis, treatment, management, and support?**

To address these key questions we adopt a big data analytics approach to processing cancer blog unstructured data.

**Big Data Analytics**

By definition, big data in healthcare refers to electronic health data sets so large and complex that they are difficult, if not impossible, to manage with traditional software and/or hardware; nor can they be managed easily with traditional or common data management tools and methods. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed. The totality of data related to patient healthcare and well-being make up “big data” in the healthcare industry. The scope includes clinical data from CPOE and clinical decision support systems (physician’s written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data); patient data in electronic patient records (EPRs); machine-generated/sensor data, such as from monitoring vital signs; social media posts, including Twitter feeds (tweets), blogs, status updates on Facebook and other platforms, and web pages; and less patient-specific
information, including emergency care data, news feeds, and articles in medical journals.

For the big data scientist, there is, amongst this vast amount and array of data, opportunity. By discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve care, save lives and lower costs. Thus, big data analytics applications in healthcare take advantage of the explosion in data to extract insights for making better informed decisions. As a research category, they are referred to as, no surprise here, big data analytics in healthcare. When big data is synthesized and analyzed - and those aforementioned associations, patterns and trends revealed - healthcare providers and other stakeholders in the healthcare delivery system can develop more thorough and insightful diagnoses and treatments, resulting, one would expect, in higher quality care at lower costs and in better outcomes overall. The potential for big data analytics in healthcare to lead to better outcomes exists across many scenarios, including: by analyzing patient characteristics and the cost and outcomes of care to identify the most clinically and cost effective treatments and offer analysis and tools, thereby influencing provider behavior; applying advanced analytics to patient profiles (e.g. segmentation and predictive modeling) to proactively identify individuals who would benefit from preventative care or lifestyle changes; broad scale disease profiling to identify predictive events and support prevention initiatives; collecting and publishing data on medical procedures, thus assisting patients in determining the care protocols or regimens that offer the best value; identifying, predicting, and minimizing fraud by implementing advanced analytic systems for fraud detection and checking the accuracy and consistency of claims; implementing claim authorizations expeditiously; and creating new revenue streams by aggregating and synthesizing patient clinical records and claims data sets to provide data and services to third parties (such as licensing data to assist pharmaceutical companies in identifying patient candidates for clinical trials). Many payers are developing and deploying mobile apps that help patients manage their care, locate providers, and improve their health. Via analytics, payers are able to monitor adherence to drug and treatment regimens and detect trends that lead to individual and population wellness benefits.

Conceptual Framework

The conceptual framework for a big data analytics project in healthcare-related social media is similar to that of a traditional health informatics or analytics project. The key difference lies in how processing the data is executed. In an ordinary structured health analytics project with routine amounts of small data the analysis can be performed with a business intelligence tool installed on a stand-alone system, such as a desktop or laptop. Because big data is by definition unwieldy, processing is broken down and executed across multiple nodes. The concept of distributed processing has existed for decades. What is relatively new is its use in analyzing very large data sets as healthcare providers start to tap into their large data repositories to gain insight for making better-informed, health-related decisions. Interestingly,
open source platforms such as Hadoop/MapReduce, available on the Cloud, have encouraged the application of big data analytics in healthcare.

While the algorithms and models are similar, the user interfaces of traditional analytics tools and those used for big data are entirely different; traditional health analytics tools have become user friendly and transparent. Big data analytics tools, on the other hand, are extremely complex, programming intensive, and require the application of a variety of skills. They have emerged in an ad hoc fashion mostly as open-source development tools and platforms, and therefore they lack the support and user-friendliness that vendor-driven proprietary tools possess.

Figure 1. An applied conceptual architecture of Big Data analytics; Source: Adapted from [34]

As Figure 1 indicates, the complexity begins with the data itself. Big data in healthcare can come from internal sources (e.g. electronic health records, clinical decision support systems, and CPOE, etc.) and external ones (government sources, laboratories, pharmacies, insurance companies & HMOs, etc.). It often appears in multiple formats (flat files, .csv, relational tables, ASCII/text, etc.), resides at multiple locations (geographic as well as in different healthcare providers’ sites), and exists in numerous legacy and other applications (transaction processing applications, databases, etc.)\textsuperscript{34}. For the purpose of big data analytics, this data has to be pooled. In the second component, the data is in a ‘raw’ state and must be processed or transformed, at which point several options are available, including data warehouse and service-oriented architecture. The data stays raw, and services are used to call, retrieve, and process the data. Via the steps of extract, transform, and load (ETL), data from these diverse sources is cleansed and readied. Depending on whether the data is structured or unstructured, several data formats can be input to the big data analytics platform. In the next component of the conceptual framework, several
decisions are made regarding the data input approach, distributed design, tool selection, and analytics models. The fourth component, on the far right, shows the four typical applications of big data analytics in healthcare. These include queries, reports, OLAP, and data mining. Visualization is an overarching theme across the four applications. Drawing from statistics, machine learning, and visualization, and other fields, a wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyze, and visualize big data in healthcare.

The most significant platform for big data analytics is the open-source distributed data processing platform Hadoop (Apache platform), which was initially developed for such routine functions as aggregating web search indexes. It belongs to the class “NoSQL” technologies-CouchDB and MongoDB are also in this class—that evolved to aggregate data in unique ways. Hadoop has the potential to process extremely large amounts of data mainly by allocating partitioned data sets to numerous servers (nodes), each of which solves different parts of the larger problem and then integrates them for the final result. Hadoop can serve the dual roles of data organizer and analytics tool. It enables enterprises to harness data that, until now, has been difficult to manage and analyze. Specifically, Hadoop makes it possible to process extremely large volumes of data with various structures or no structure at all. It has been confirmed that very large data sets with complex structures are difficult to process using traditional methods and tools. The complex process includes capture, storage, formatting, extraction, curation, integration, analysis, and visualization. Therefore, more robust and scalable platform tools are needed for managing big data. Cloudera (http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/) provides the scalable, flexible, and integrated platform that makes it easy to manage rapidly increasing volumes and varieties of data in the domain. The Cloudera platform can be deployed to manage Apache Hadoop and related projects, manipulate and analyze data, and keep that data secure and private.

Methods
In keeping with our research objectives, we show the potential of Hadoop MapReduce to extract patterns and insight on cancer diagnoses from cancer blogs. We apply a series of algorithms to gain insight into the content and develop a vocabulary and taxonomy of keywords based on existing medical nomenclature. Some key questions can provide insight into diseases (cancer), treatments (e.g. alternative medicine, therapy), and support links. What are the most common issues patients have (bloggers/responses)? What cancer types (conditions) are most discussed and why? What therapies and treatments are being discussed? What medical and nonmedical information is provided? Which blogs and bloggers provide relevant and accurate information? What are the major motivations for the postings (comments)? Who posts—doctors, nurses, or patients? What are the emerging trends in disease (symptoms), treatment and therapy (e.g. alternative medicine), support systems, and information sources (links, clinical trials)?
Data Collection
Data was collected from the website [http://www.thecancerblog.com](http://www.thecancerblog.com). Most of this data was unstructured, such as responses and comments. Figure 8 shows a sample posting. In order to use the text mining algorithm, we first created two input files: text.dat and dates.dat. Each line in the text.dat file stores the blog text (the blog entry and the comments posted to it), and each line in the dates.dat file stores the corresponding blog creation dates. Initially, all blogs were downloaded directly from the source website and stored in html files. Since the information required by the architecture is mixed with numerous advertisements, hyperlinks, and JavaScripts, extraction programs in Java were then used to extract the blog texts and creation dates. One important point to keep in mind is that because the content is collected from the web, it includes extraneous details. The application enabled semantic searches by the extraction and organization of concepts and relationships.

Analysis and Results
Hadoop MapReduce is a software framework for writing applications that process—in a reliable, fault-tolerant way—vast amounts of data in parallel and on large clusters (# of nodes) of servers. A MapReduce job typically splits the input data set into independent chunks, which are processed by the “map tasks” in a completely parallel manner. The framework sorts the outputs of the maps. These are then input to the “reduce tasks”. Typically, both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them, and re-executing failed tasks. Usually, the compute nodes and the storage nodes are the same. That is, the MapReduce framework and the Hadoop Distributed File System are running on the same set of nodes. This configuration allows the framework to effectively schedule tasks on the nodes where data is already present, resulting in very high aggregate capacity across the cluster. The MapReduce framework consists of one master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for scheduling the jobs’ component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed, by the master.

At a minimum, applications specify the input/output locations and supply map and reduce functions via implementations of appropriate interfaces and/or abstract classes. These and other job parameters comprise the job configuration. The Hadoop job client then submits the job (jar/executable, etc.) and configuration to the JobTracker, which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, and providing status and diagnostic information to the jobclient. The Cloudera Hadoop MapReduce platform was deployed for the entire project.

In order to extract the key topics discussed in cancer blogs, the 1606 reviews in the cancer blog data set were analyzed by applying the typical clustering algorithm. The purpose was to elicit word groupings, calculate word frequency, and identify word-pair associations. Standard data preparation was conducted first. Python NLTK library was used to eliminate punctuations, numbers, and standard stop words. In order to enhance the quality of the words in clustering, additional stop words such as “feel”, “would”, “take,” and others were incorporated. The
Snowball stemming algorithm was used to combine different words with similar meanings into a single word. For instance, “fishing” and “fishes” were stemmed into “fish.” Next, with the 1606 text files with stemmed words as the input, K-means clustering, word count, and word co-occurrence algorithms in the Cloudera Hadoop MapReduce platform were applied. The blog post text files were converted into sequence files, and the TF-IDF index was then constructed. The Apache Mahout library was then utilized to perform clustering analysis and output the results from the HDFS to the local machine. Figure 2 shows the top ten keyword clusters, with thirty terms in each cluster.

![Figure 2. Keyword clusters](image)

In figure 2, the 333 blog posts in Cluster 1 mention thirty terms, including “prostate cancer”, “hormone”, “drug” and “risk”, indicating the posts were associated with medication and risk. This cluster could be labeled as “prostate cancer.” Cluster 2, with 248 postings, is dominated by breast cancer terms, including “breast cancer survivor,” “diagnose breast cancer,” “pink,” (international symbol of breast cancer awareness), and so on. Other clusters are labeled similarly. The third cluster relates to how patients feel generally (“feeling”). The fourth cluster (named “Treatment”), with 200 postings, encompasses treatment and surgery, with terms such as “transplant,” “home,” and “treatment.” Cluster 10, with sixty postings, indicates the discussion is primarily about the Susan G. Komen organization. Therefore, we can label it as “Komen.” Meanwhile, Cluster 5 pertains to “chemo” & related effects;
Cluster 6 indicates “survivors”; Cluster 7 describes “therapy”; Cluster 8 appears to reflect “HPV”; and Cluster 9 focuses on ‘side effect’.

The input data for this experiment was the original cancer blog text file, which was uploaded to the HDFS file system. Here we run a word count job to learn how many times a particular word has been used. We approach this task by leveraging the power of Cloudera platform services and then assembling a Java program to count the words in the given data set. We used the Cloudera Hadoop platform to perform word count analysis using built-in Java source code provided by Cloudera. In this experiment, HDFS commands were used to navigate to a folder that contained three different java files: WordCount.java, SumReducer.java, and WordMapper.java. Another command was then used to compile the three Java files into a single JAR file. This JAR file was a MapReduce job, which performed a word count of the words used in all of the text files contained in the folder “Cancer-blog”.

Map and reducing data can be based on a variety of criteria. A common example is the Java WordCount class. As the name suggests, WordCount maps (extracts) the words in the input and reduces (summarizes) the results with a count of the number of instances each word is used. WordCount reads text files and counts how often words occur. The input is text files, as is the output, and each line contains a word and the count of how often it occurs, separated by a tab. Each mapper takes a line as input and breaks it into words. It then generates a key/value pair of the word and 1. Each reducer adds the counts for each word and emits a single key/value with the word and the sum. As an optimization, the reducer is also used as a combiner on the map outputs. This reduces the amount of data sent across the network by combining each word into a single record. Figure 3 shows the result of the word count algorithm to cancer blogs. The larger the word, the higher its frequency. As the word cloud indicates, the top issues are diagnosis and treatment of breast cancer.

Figure 3. Word cloud
This process required Hadoop to go about the usual steps of creating a sequence file and then TF-IDF vectors. TF-IDF stands for term frequency-inverse document frequency and reflects how important a word is to a document or corpus. The application of TF-IDF vectors helps to determine what words in a corpus of documents might be more favorable to use in a query. As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if the word were to appear in a query, the document could be of interest to the user. This simple algorithm efficiently categorizes relevant words that can enhance query retrieval.

Once TF-IDF vectors were created, those vector files were used to find similar files for each file in the file store. Distances between the vectors helps the algorithm to determine how closely two documents are related in terms of content. The last step was to create a report of the final-row-similarity job and the distances between document-pairs give you fairly good understanding as to the distance of similarity between the two documents. Each key value in the final output represents an individual input text file. Figure 4 shows the similarity of different blog postings. For example, Key 0 shows the 948th posting is similar to the 433rd, 801th, and 1525th postings - all of which relate to cancer mitigation.

The next analysis applied the word co-occurrence algorithm. The term co-occurrence implies concurrent occurrence of two terms from a text corpus alongside each other in a certain order, indicating semantic proximity. Therefore, co-occurrence can be adopted to quantify semantic relations between words. Here, “events” are the individual words found in the text, and they track other words that occur within a “window,” a position relative to the target word.

For this algorithm, we used Cloudera Hadoop platform to perform word co-occurrence analysis using built-in Java source code provided by Cloudera.
code for word co-occurrence was compiled to create a JAR file, which would serve to provide the instructions for our data processing using Hadoop and MapReduce. Before we could produce an executable JAR file, we needed to compile the source code, which included our mapper and reducer scripts. After compiling the source code, we converted the java file into an executable JAR file, and uploaded that JAR file to the Hadoop File System. We then ran the MapReduce job, which read our unstructured text input file and made a list of all the paired words that were used within the input file. For each word in the input file, the list specified every word-pair, which included that word and listed how many times that word-pair appeared. The top 20 co-concurrent words are displayed in Figure 5.

![Figure 5. Word co-occurrence](image)

Our purpose here (figure 5) is to identify what combination of binary words adds insight to the analysis. We streamlined the number of co-occurring words because the overall output included more than 20 thousand word pairs. Therefore, via careful pruning combinations, we identified and eliminated such terms as ‘one a day,’ ‘cancer and nation,’ ‘would and like.’ We learned, for instance, that ‘breast’ and ‘cancer’ are paired at least 3409 times. This number indicates ‘breast cancer’ dominates blog discussions. ‘Skin cancer’ occurs 188 times, ‘prostate cancer’ occurs 408 times, and ‘lung cancer’ occurs 304 times.

Using the Cloudera platform with Hadoop and MapReduce, we then performed a classification analysis, employing a naïve-Bayes classification model on the clustered cancer blog data files. We wanted to discern whether there there are
blogs that belong in clusters other than the ones in which analytics placed them. The output of clustering process then serves as the input to the classification process. This process checks the accuracy of automated clustering analysis while identifying misplaced blog postings. We found some of the cancer blog postings were very closely related to each other, while others were entirely unrelated but had nevertheless been placed in the same cluster.

The purpose of classification analysis is to create and train a predictive model that can read the contents of each document and correctly sort documents into the appropriate cancer blog grouping or cancer blog cluster. With the uploaded clustered cancer blog data, we created a sequence file, a vector file, and a naïve-Bayes classifier. The data is divided the data into two groups - training and testing - based on the specified proportions. Seventy percent of the data was assigned to the training portion. 30% was assigned to the testing portion, resulting in a confusion matrix.

As shown below in Figures 6 and 7, the model was able to accurately predict the correct cancer blog cluster for each document and produced the model with 99.82% accuracy. Our training portion contained a total of 1,102 documents, 1,100 of which were classified correctly. The high percentage accuracy of our model is very promising and indicates that this model could be used to classify new documents in the future. If documents are currently being classified and sorted manually, this model represents a potentially significant reduction in man-hours and, by extension, large cost reductions.

Figure 6. Confusion matrix for classification analysis
Summary

Correctly Classified Instances: 99.82%
Incorrectly Classified Instances: 0.18%
Total Classified Instances: 1102

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Actual Positive</th>
<th>Actual Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Positive</td>
<td>1098</td>
<td>4</td>
</tr>
<tr>
<td>Predicted Negative</td>
<td>5</td>
<td>1175</td>
</tr>
</tbody>
</table>

Figure 7. Summary results of classification analysis

Figure 8 is a cross section of the blog content. Figure 9 displays the number of blog postings in the study.
In summary, we applied the big data analytics architecture of Hadoop MapReduce via the Cloudera platform, to the analysis of cancer blog content. By applying a number of algorithms, we gained insight into the blog content. The study identifies, for instance, the most discussed topics as well as associations that relate to key phenomena.

Scope and Limitations
This study focuses on a single category of blogs relating to healthcare—cancer blogs. Obviously, a vast number of blogs exist. Indeed, there is a wide array of topics not covered in our analysis, and so the potential for generalizability is limited. Also, the study analyzes the blog content existing at a certain point in time. Given the rapid changes in content and the number of blogs, the health blogosphere will continue to evolve in ways that, naturally, will vary from this snapshot. Finally, the data we analyzed should be characterized as self-reported and subjective. Thus, it can (and should) be scrutinized in terms of accuracy and veracity. Indeed, the quality and validity of blog postings in general is questionable. A more extensive analysis can encompass more websites and a longitudinal study of the content.

Technology is a limitation, especially as it is advancing rapidly. Future research may utilize more sophisticated and advanced platforms and algorithms, such as SPARK, and convolutional neural networks. Also, a big data analytics
platform for healthcare social media must support the key functions necessary for processing the data. The criteria for platform evaluation may include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance. In addition, while most platforms currently available are open source, the typical advantages and limitations of open source platforms apply. To succeed, big data analytics in social media needs to be packaged so it is menu-driven, user-friendly, and transparent. Real-time big data analytics is a key requirement in healthcare. The lag between data collection and processing has to be addressed. The dynamic availability of numerous analytics algorithms, models and methods in a pull-down type of menu is also necessary for large-scale adoption. The important managerial issues of ownership, governance, and standards have to be considered. And woven through these issues are those of continuous data acquisition and data cleansing. Social media data is rarely standardized, often fragmented, or generated in legacy IT systems with incompatible formats. This great challenge must also be addressed.

Conclusions
Big data analytics has the potential to transform the way practitioners and researchers gain insight from health social media, especially those in free text, unstructured form. Considering the volume and variety of social media data, and the need for scalability and large scale processing in real-time, we should in the future see rapid, widespread implementation and use of big data analytics across the healthcare social media content. To that end, the several challenges highlighted above must be addressed. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in such healthcare-related social media as blogs are at an early stage, but rapid acceleration is possible with the advancements in models, tools, and technologies.

References


