Exploring Big Data Analytic Approaches to Cancer Blog Text Analysis

Viju Raghupathi, Koppelman School of Business, Brooklyn College of the City University of New York, Brooklyn, USA
Yulu Zhou, Gabelli School of Business, Fordham University, New York, USA
Wullianallur Raghupathi, Gabelli School of Business, Fordham University, New York, USA

ABSTRACT

In this article, the authors explore the potential of a big data analytics approach to unstructured text analytics of cancer blogs. The application is developed using Cloudera platform’s Hadoop MapReduce framework. It uses several text analytics algorithms, including word count, word association, clustering, and classification, to identify and analyze the patterns and keywords in cancer blog postings. This article establishes an exploratory approach to involving big data analytics methods in developing text analytics applications for the analysis of cancer blogs. Additional insights are extracted through various means, including the development of categories or keywords contained in the blogs, the development of a taxonomy, and the examination of relationships among the categories. The application has the potential for generalizability and implementation with health content in other blogs and social media. It can provide insight and decision support for cancer management and facilitate efficient and relevant searches for information related to cancer.

KEYWORDS

Big Data Analytics, Cancer Blog, Classification, Cloudera, Clustering, Hadoop, Mapreduce, Row Similarity, Text Analytics, TF-IDF, Word Co-Occurrence, Word Count

1. INTRODUCTION

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 (Chen et al., 2015; Greaves et al., 2013). 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, web search logs, among others (Katsuki et al., 2015; Mazzocut et al., 2016; Surian et al., 2016). 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 (Agarwal et al., 2016; Chan et al., 2013; Chen et al., 2015; Yom-Tov et al., 2014). 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 (Miller & Pole, 2010). Hospitals and other providers use these tools as “gateways” to the communities (Hardy, 2012; Kotenko, 2013; White, 2015). 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 (Bian et al., 2012; Konkel, 2013). Spasic et al. (2014) discuss how so-called 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., (Kim, 2009; Lin et al., 2011; Moen et al., 2016; Spasic et al., 2014; Wright et al., 2010; Zhu et al., 2013).

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 (Mulins et al., 2006). 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) (Spangler & Kreulen, 2007). Text analytics is typically used to identify patterns and trends in the unstructured data (Popowich, 2005). 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 (Rauber, 2004).

With regard to health social media analytics, several papers have examined the potential. Surian et al. (2016), for instance, analyzed 285,417 Twitter posts, also known as tweets, about HPV vaccines. They studied the tweets of some 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 (Surian et al., 2016).” 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.

Jung et al. (2016) focused on identifying the quality of hospital service automatically using online communities. 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. Lardon et al. (2015) examined the potential for post marketing safety surveillance from patient experiences with drugs reported
Improving the Decision-Making Process in a Hospital Environment With New Interactive Visualization Methods

Implementation of a Referent Tracking System
www.igi-global.com/article/implementation-referent-tracking-system/2215?camid=4v1a