Chapter 90
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

Yilu 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.

DOI: 10.4018/978-1-6684-3662-2.ch090

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).