Textual and Informational Characteristics of Drug-Related Content on Three Kinds of Websites: Drug Review Website, Discussion Board and Hospital Information Portal

Lorraine Goeuriot, Nanyang Technological University, Singapore
Jin-Cheon Na, Nanyang Technological University, Singapore
Wai Yan Min Kyaing, Nanyang Technological University, Singapore
Christopher S.G. Khoo, Nanyang Technological University, Singapore
Yin-Leng Theng, Nanyang Technological University, Singapore
Yun-Ke Chang, Nanyang Technological University, Singapore
Schubert Foo, Nanyang Technological University, Singapore

ABSTRACT

This paper develops techniques to mine medical and health-related social media content for information and opinions about diseases and treatments. This would allow users to benefit from the numerous postings of people’s experiences scattered all over the Web. This paper reports the results of a textual analysis of user-generated content on drug-related online sites. The aim is to determine what kinds of content can be expected on these sites, from linguistic and information points-of-view. User postings were harvested from two websites carrying different kinds of user-generated content and compared to information on the same drugs from a hospital information portal. The corpus was analyzed to identify what kinds of drugs were often reviewed, the vocabulary and medical concepts used, and the textual characteristics such as length of postings, sentence length, and part-of-speech distribution. Although drug-related user-generated content is very different from editorial content, the results of this project show that it provides useful information on many drugs. From a linguistic point-of-view, user-generated content is simpler in language and more informal in style than editorial content, but still contains useful health information that tends to be patient-oriented.

Keywords: Content Analysis, Discussion Forums, Drug Reviews, Medical and Health Information, Opinion Mining, Social Media, User-Generated Content

DOI: 10.4018/joci.2011070103
1. INTRODUCTION

The quantity of online health and medical information is constantly increasing. Health related websites have become one of the most important public information sources on health, disease and treatment (Sarasohn-Kahn, 2008). Web 2.0 technologies have given rise to a host of social media sites, such as discussion forums, blogs, and user-review sites, which allow users to contribute their own content. The emergence of such new data in large quantities has given rise to questions about information searching and quality of information. Finding relevant, useful and trustworthy information on the Web is a real challenge. One of the attractions of social media is the support for sharing information among people having similar experiences and situations.

Previous studies have shown that this user-generated content is useful from different points of view. First, users are often looking for stories from “patients like them” on the Internet, which they cannot always find among their friends and family (Sarasohn-Kahn, 2008). Moreover, studies investigating the impact of social media on patients have shown that for some diseases and health problems, online community support can have a positive effect (e.g., Jaloba, 2009; Schraefel et al., 2009). Because of its novelty as well as quality and trustworthiness issues, user-generated content is underexploited. It needs to be further studied, understood and then leveraged in designing new online tools and applications.

The overall goal of our project is to develop techniques to mine medical and health-related social media content for information and opinions about diseases and treatments. In a previous paper (Goeuriot et al., 2011), we reported preliminary results of a textual analysis of user-generated content on drug review websites. In this paper, we extend the analysis to three different drug information sources: drug review websites, discussion boards and hospital portals. We evaluate the quality of the websites and their content in the light of opinion mining on drugs. In this context, we define quality according to three criteria: (1) Websites must cover a wide range of drugs; (2) They must contain texts that are both informative (i.e. containing medical information) and affective (i.e. containing opinions); (3) The texts should be syntactically and lexically correct. However, in this study, the validity of social media content is not investigated since it requires in-depth manual analysis with medical domain knowledge.

2. RELATED WORKS

Health information on the Internet is a highly discussed topic, raising both concern and enthusiasm. Recently, health on social media, so-called Health 2.0, is attracting much interest. According to Sarasohn-Kahn (2008, p. 2), it can be defined as: “the use of social software and its ability to promote collaboration between patients, their caregivers, medical professionals and other stakeholders in health”. Social software refers to blogs, discussion forums, podcast, social networks and wikis. We will focus the literature review on three kinds of studies: social media users, user-generated content, and use of this content and its applications.

2.1. Health-Related Social Media Users

According to Nettleton et al. (2005), the increasing use of the Web for health matters raises three kinds of reaction: concern for physicians, enthusiasm for sociologists, and a contingent and embedded response.

The California Healthcare Foundation reported that, according to a 2008 survey in the
Classifying Web Pages by Aimed Nation Using Machine Learning
www.igi-global.com/article/classifying-web-pages-by-aimed-nation-using-machine-learning/165390?camid=4v1a

Theoretical Derivations and Application Issues
E. Parsopoulos Konstantinos and N. Vrahatis Michael (2010). *Particle Swarm Optimization and Intelligence: Advances and Applications* (pp. 42-87).
www.igi-global.com/chapter/theoretical-derivations-application-issues/40630?camid=4v1a