Chapter 98
Mining User-Generated Content for Social Research and Other Applications

Rafael E. Banchs
Institute for Infocomm Research, Singapore

Carlos G. Rodríguez Penagos
Barcelona Media Innovation Centre, Spain

ABSTRACT

User-generated content is currently becoming a valuable means for sensing and measuring real world variables and parameters that are of interest to several actors in the society: politicians, government departments, security agencies, marketing researchers, service providers, etc. In response to this new scenario, large research efforts are being invested in the so-called "social media" phenomenon by a wide spectrum of institutions and organizations around the world, with many different objectives and a diverse scope of fields and disciplines. As a consequence, new technologies and applications are currently emerging on the grounds of human participation, interaction, and behavior on the Internet.

The main objective of this chapter is to present a general overview of the most relevant applications of text mining and natural language processing technologies evolving and emerging around the Web 2.0 phenomenon (such as automatic categorization, document summarization, question answering, dialogue management, opinion mining, sentiment analysis, outlier identification, misbehavior detection, and social estimation and forecasting) along with the main challenges and new research opportunities that are directly and indirectly derived from them.

DOI: 10.4018/978-1-4666-3886-0.ch098

INTRODUCTION

Internet has changed human communications in several different ways, but perhaps one of the most prominent changes has come hand-in-hand with the so-called Web 2.0. Within the scenario of these second generation of Web technologies, information is generated, delivered and consumed by "end" users of traditional mass media communication: the general public. According to this view, and as several Web 2.0 applications currently demonstrate,[1] the ability to broadcast has become available, theoretically, to almost everybody. This constitutes one major milestone since the introduction of mass media communication experimented by our modern information society, with deep repercussions in culture, sociology and economics (Ala-Mutka, et al., 2009). Consequently, traditional technologies are evolving and new technologies and applications are emerging from, and for, the Web 2.0. Some examples of these are opinion mining (Funk, et al., 2008), sentiment analysis (Pang & Lee, 2008), question answering (Chali, 2009), user profiling (Kontostathis, et al., 2009), recommender systems (Ricci & Werthner, 2006), behavioral marketing (Berkman, 2008), and social forecasting (Durant & Smith, 2007), among others.

One of the most important issues regarding the Web 2.0 phenomenon is that most of the user interactions, as well as generated contents, involve human language; so the Web 2.0 era demands natural language processing technologies more than ever. Main types of user-generated text range from formal journalistic and/or biographic contents, such as in the case of blogs; to shorter and more informal contents, such as discussion forums and consumer reviews, opinions and recommendations; and, in the micro-blog extreme, to very short messages known as "tweets." All these impose particular requirements of speed and efficiency on traditional natural language applications, as well as accuracy, scalability and robustness requirements that are difficult to tackle with current available technologies. Additionally, when considering natural language processing techniques, new communication styles, and varieties of language usage must be taken into account too. The extended use of non-standard practices such as emoticons (character sequences denoting emotions or gestures) and chatspeak (special spelling and terminology associated with informal social media exchanges), as well as their corresponding context-dependent protocols, are generating new communication and language "standards" that have to be tackled by analysis applications focused on user-generated content.

At this moment, it is still not possible to fully foresee the implications and consequences of massive user-generated content analysis in terms of both scientific and commercial exploitation. This is mainly because of the complex and multidimensional nature of the Web 2.0 phenomenon itself. Indeed, practical experience demonstrates that, although many Web 2.0 websites have been growing at a very fast pace during the last few years, only few of them have actually been able to develop successful business models from their corresponding virtual communities. Nevertheless, what it is actually possible to foresee are some trends emerging from this new framework of human interaction and communication. The main objective of this chapter is to present a general overview of the most relevant applications of text mining and natural language processing technologies currently emerging around the Web 2.0 phenomenon, along with the main challenges and new research opportunities that are derived from them.

The chapter is structured as follows. First, a background section providing the main definitions and general discussions on social media and user-generated content analysis is presented. This section covers some fundamental issues regarding the Web 2.0, social media, and natural language processing technologies, which should provide necessary background for the following sections. Second, a section on technical challenges is pre-

Innovation in the Time of Pandemic: Insights from a Survey of Malaysian Small and Medium Enterprises (SMEs)
Mohammed Alnajjar, Abdelhak Senadjki, Au Yong Hui Neeand Samuel Ogbeibu (2025). *International Journal of SME Research and Innovation (pp. 1-21).*
www.irma-international.org/article/innovation-in-the-time-of-pandemic/368040

Innovation in the Time of Pandemic: Insights from a Survey of Malaysian Small and Medium Enterprises (SMEs)
Mohammed Alnajjar, Abdelhak Senadjki, Au Yong Hui Neeand Samuel Ogbeibu (2025). *International Journal of SME Research and Innovation (pp. 1-21).*
www.irma-international.org/article/innovation-in-the-time-of-pandemic/368040

Innovation in the Time of Pandemic: Insights from a Survey of Malaysian Small and Medium Enterprises (SMEs)
Mohammed Alnajjar, Abdelhak Senadjki, Au Yong Hui Neeand Samuel Ogbeibu (2025). *International Journal of SME Research and Innovation (pp. 1-21).*
www.irma-international.org/article/innovation-in-the-time-of-pandemic/368040