Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective

E. Teso[a], M. Olmedilla[b], M.R. Martínez-Torres[c,*], S.L. Toral[c]

[a] *Liverpool Business School, Liverpool John Moores University, Redmonds Building Brownlow Hill, Liverpool L3 5UG, United Kingdom*
[b] *Léonard de Vinci Pôle Universitaire, Research Center, 92 916 Paris La Défense, France*
[c] *Facultad de Turismo y Finanzas, University of Seville, Avda. San Francisco Javier, s/n, 41018 Sevilla, Spain*

ARTICLE INFO

ABSTRACT

The emergence of online user-generated content has raised numerous questions about discourse gender differences as compared to face-to-face interactions. The intended gender-free equality of Internet has been challenged by numerous studies, and significant differences have been found in online communications. This paper proposes the application of text mining techniques to online gender discourse through the analysis of shared reviews in electronic word-of-mouth communities (eWOM), which is a form of user-generated content. More specifically, linguistic issues, sentiment analysis and content analysis were applied to online reviews from a gender perspective. The methodological approach includes gathering online reviews, pre-processing collected reviews and a statistical analysis of documents features to extract the differences between male and female discourses in a specific product category. Findings reveal not only the discourse differences between women and men but also their different preferences and the feasibility of predicting gender using a set of frequent key terms. These findings are interesting both for retailers so they can adapt their offer to the gender of customers, and for online recommender systems, as the proposed methodology can be used to predict the gender of users in those cases where the gender is not explicitly stated.

1. Introduction

Over the last few decades, language has become a very important aspect of gender studies. Language is an inherently social phenomenon and can provide an insight into how men and women approach their social worlds (Newman et al., 2008). In fact, many studies have focused on analysing to what extent men and women use language differently (Brownlow et al., 2003; Colley et al., 2004). There is evidence to indicate meaningful differences in men's and women's language, not only differences in pronunciation and intonation, but also in vocabulary and syntax (Xia, 2013). For instance, women are likely to use more intensive adverbs, more conjunctions, and more modal auxiliary verbs, while men swear more, use longer words, use more articles, and use more references to location (Mehl and Pennebaker, 2003). Similarly in social interactions, gender differences regarding chosen topics have been reported. Women are generally more interested in talking about family affairs, such as the education of children, clothes, cooking, fashion, etc. whereas men are more likely to choose topics such as politics, economy, stock and shares, sports and current news (Xia, 2013).The words people use in their daily lives can reveal important

aspects not only about their personalities but also about their preferences or choices (Pennebaker et al., 2003). This information can have a business value, as many retailers are interested in knowing gender differences when customers make purchasing decisions or in online recommender systems.

A traditional limitation of gender discourse studies is that large samples are required to obtain broad conclusions about how men and women differ in their language use across settings. Today, the emergence of user-generated content is alleviating this problem, as the Internet can provide many texts on a wide variety of themes that can be easily identified as belonging to men or women. For instance, opinion websites or electronic word-of-mouth (eWOM) communities make a vast amount of information about people's preferences accessible. Gender information is something web users are not usually too reluctant to provide on online communities. A second limitation relates to the methodological perspective when analysing discourse differences. Qualitative analyses are difficult to carry out when the data sample is large. In contrast, quantitative analyses rely on natural language processing, which is based on the idea that features of language or word use can be counted and statistically analysed. This paper overcomes

these two limitations by using data from a popular eWOM community. eWOM is the most pervasive form of user-generated business discourse (Vásquez, 2015). It involves positive or negative statements made by customers about a product or a company which is made available to many people via the Internet (Hennig-Thurau et al., 2004), so it can be considered a specific type of user-generated content focused on products or companies (Wang and Rodgers, 2010). This study focuses on eWOM communication in a consumer opinion website, ciao.co.uk, which enables consumers to post reviews of products and services or simply view other consumers' experiences. More specifically, the study focuses on the category of books. The overall objective is the application of text mining techniques, such as linguistic dimensions analysis, sentiment analysis and content analysis to find if there are significant differences in the discourse of men and women when reviewing and commenting books. The application of the two first techniques emphasizes the gender differences regarding the linguistic dimension (the use of pronouns, swear words, auxiliary verbs …) and emotions (positive and negative emotions, anger, anxiety …). Content analysis is a method for interpreting the meaning of texts and quantifying the frequency of those interpretations (Dumay and Cai, 2014), and it can be used to determine the thematic preferences, or even for gender prediction in those reviews where the gender of the author is not provided. Unlike previous discourse gender studies, the use of computational techniques for collecting, pre-processing and analysing documents can automate the whole process, so a large sample of documents can be considered. By answering this objective, this paper aims to contribute to the general debate of whether the language of online communication is gender free or whether it is influenced by patterns of male dominance which have traditionally been observed in face-to-face communication. Hence, this study analyses if the potential degree of invisibility and anonymity afforded by the Internet (Hollenbaugh and Everett, 2013) is present in a eWOM website by analysing online consumer reviews. Note that we refer to socially constructed gender not biological sex.

The paper is structured as follows: next section briefly reviews the literature in the field of text analysis related to gender discourse studies; Section 3 introduces the research questions; Section 4 describes the research design, including data collection and proposed methodology; Section 5 describes the findings and discusses their meaning and implications, and finally, Section 6 details the conclusions.

2. Related work

2.1. Discourse gender differences

The study of language and gender has increasingly become the study of gender and discourse. Discourse has been defined as language in context (Bucholtz, 2003). Context refers to the language as it is used in social situations and interactions. Most narrative researchers assume that language is, by definition, contextual so phrases or entire texts must be considered within the context of the speaker and the relationship between the speaker and the audience (Pennebaker et al., 2003). Differences in the way men and women use language have long been of interest in the study of discourse. Previous research on gender differences in language use suggests that men tend to use language more for the instrumental purpose of conveying information while women are more likely to use verbal interaction for social purposes with verbal communication serving as an end in itself (Newman et al., 2008). Women are more likely to ask questions in daily conversations while men are more direct and use more commands telling people to do something. Women have also been found to use longer sentences while men use more words overall and take more speaking "turns" in conversation (Mulac and Lundell, 1994; Newman et al., 2008). There are some other aspects where there is no clear consensus. For instance, some studies report that women are more likely to use first-person singular pronouns (Mehl and Pennebaker, 2003) although the word "I" connotes individualism or selfishness, which fits the male stereotype

better than the female stereotype (Newman et al., 2008). Regarding emotions, most studies report that women refer to emotion more often than men (Thomson and Murachver, 2001). According to Mehl and Pennebaker (2003), women use more references to positive emotion whereas men use more to negative emotion words in their daily language.

The advent of the Internet and Web 2.0 has opened new opportunities for analysing the relationship between gender and discourse. Computer-mediated communication (CMC) is essentially a different context with its own characteristics (Herring and Stoerger, 2013). Electronic discourse shows elements of both written language and speech. It is sent and received so rapidly that it nears the interactivity of speech, and conversations can emerge. Electronic messages are less censored and more informal than writing intended for hard copy, and tend to mimic the paralinguistic features of spoken conversation (Thomson and Murachver, 2001). The emergence of user-generated content and virtual communities have broaden the scope of gender studies (Zhang et al., 2013). One of the most significant topics in the literature focuses on the anonymity of the internet that makes gender of online communications invisible or irrelevant (Herring, 2003), as users share information hidden behind an alias, where the gender is not usually evident. Some studies suggest similarities in the way in which men and women use and interact within online support groups. For instance, Thomson and Murachver (2001) conclude that men and women are equally likely to ask questions and offer compliments, apologies, and opinions in email communication. However, although gender anonymity would supposedly allow women and men to participate equally on online communications as opposed to the male dominance traditionally observed in face-to-face communication (Herring, 2003), there is a large body of evidence that contradicts this claim (Dwight, 2004; Herring and Stoerger, 2013). In this line, some studies reported gendered styles of communicating (Blank and Adams-Blodnieks, 2007; Burri et al., 2006). An extensive review about gender differences in computer-mediated communication can be found in Mo et al. (2009).

It has also been highlighted the relevant role of gender in virtual communities, both in communication and e-commerce transactions (Ulbrich et al., 2011). Women are more likely to use the Internet to give and receive social support and their e-commerce transactions are more emotional, while men tend to be more pragmatic in their communication style and in their e-commerce transactions (Fan and Miao, 2012). Previous studies have also focused on gender differences when expressing positive or negative orientation, or emotions such as happiness, love, and life satisfaction (Newman et al., 2008). However, most studies typically consider face-to-face communication rather than communication through social media, despite the importance that electronic communication has today. This is because many researchers still assume that text-based online communications could not transmit socio-emotional content the same way as face-to-face communications (Sproull and Kiesler, 1986). However, later analyses showed that text-based online communications can support socio-emotional and relational communications when users feel they are connected to their online communities (Walther, 1992; Zhang et al., 2013). This paper contributes to the debate about gender neutrality of the web by analysing gender differences in online communications from a gender perspective and by comparing these differences with those reported for face-to-face communication.

2.2. Thematic preferences from a gender perspective

Opinion and eWOM websites provide an excellent resource for companies to publicly access shared opinions and understand consumers' preferences. Prior studies have applied various opinion mining methods using natural language processing or text mining techniques (Xiao et al., 2016; Yan et al., 2015). However, only a few number of studies considered the variable gender as an input for determining such