Application of Text Mining techniques to the Analysis of Discourse in eWOM Communications from a Gender Perspective

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.

Keywords electronic word-of-mouth; eWOM; user-generated content; discourse analysis; sentiment analysis; content analysis.

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 & 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 & 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 & 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 & 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 (Newman et al., 2008; Mulac & Lundell, 1994). 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 & 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 & 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 & 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, 2008), 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 & Stoerger, 2013). In this line, some studies reported gendered styles of communicating (Burri et al., 2006; Blank & Adams-Blodnieks, 2007). 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 & 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 (Yan et al., 2015; Xiao et al., 2016). However, only a few number of studies considered the variable gender as an input for determining such consumer preferences. Kim et al. (2007) focused on online travel information and obtained some significant items that could be associated to females and males. More specifically, women were more

interested in terms like ‘‘entertainment,’’ ‘‘local information,’’ ‘‘restaurant,’’ and ‘‘map’’ while men were more likely interested on information related to ‘‘flight’’, ‘‘accommodation’’, ‘‘rental car’’ and ‘‘weather’’. It can be noticed that generally, women are more interested about more subjective features than men, more focused about topics that can be quantitatively measured in terms of price or temperature. This conclusion is in line with the study by Park et al. (2009), which concluded that females are likely to read customer reviews more when shopping for experience goods (subject to more subjective standards derived from the experience with the product) than when shopping for search goods (subject to more objective standards). A tendency for men and women to prefer designs produced by people of the same gender was postulated by Moss et al (2007).

In the specific context of books, gender stereotypes have been studied regarding children’s book preferences (Mohr et al., 2006). Findings of this study reveals that boys prefer books about male characters while girls prefer stories with female protagonists. Girls are also more enthusiastic about stories highlighting family, friendships, and home life. In contrast, boys prefer nonfiction stories, particularly sports, science, and history information. This study also reveals that gender stereotypes are not inherited but more culturally influenced.

2.3. Methodological approaches to the analysis of discourse

One important limitation of previous studies is the size of the sample. Before the advent of the Internet, large samples were difficult to collect and the hand coding of features also limited the scope of studies (Newman et al., 2008). With the emergence of user-generated content, this limitation can be overcome, as there are many websites where people share

information. Hand coding can also be replaced by natural language processing techniques, which are able to computationally collect and process hundreds or thousands of documents. To do that, the computer programme must browse the website or part of it following its hyperlink structure while downloading the information of interest (Martinez-Torres, 2014). A computer programme that performs this double objective is called a crawler, and it must be customised for each case, as every website is structured and programmed in very different ways (Martinez-Torres & Olmedilla, 2016). Once the reviews are recorded and available, there a number of text mining techniques that can be applied depending on the objective of the study.

There are three text mining techniques that have been applied to discourse analysis. They rely on the extraction of relevant terms or keywords (obtained from the text itself or from dictionaries) and the calculation of features associated to these terms such as the Term frequency (TF) or the TF-IDF (Term-Frequency-Inverse Document Frequency). Then, statistical analyses or machine learning techniques can be applied to these features in order to obtain some interpretation of collected documents (Martínez-Torres & Díaz-Fernández, 2014).

The first technique is sentiment analysis, which refers to the detection and classification of the sentiment expressed by an opinion holder. Sentiment analysis, also known as “opinion mining”, summarizes large amounts of data to identify sentiment polarity (positive, negative or neutral) or the prevailing emotions (anger, sadness, trust …) to make the online generated content easier to process and understand (Khan et al., 2014; Kim et al., 2017). There are two main approaches to extracting sentiments automatically (Pang and Lee, 2008). The first is the lexical-based approach, which involves calculating

the semantic orientation for an online review by determining whether the sentences express positive or negative feelings (Zhang et al., 2013; Martinez-Torres et al., 2013). This approach is based on a dictionary of predefined list of opinion words, where each word is associated with a specific sentiment. The text's overall semantic orientation or prevailing emotion is then determined by aggregating the individual word scores, as retrieved from the sentiment lexicon (Hogenboom et al, 2014). Examples of such dictionaries are LIWC, AFIN-111, etc (González-Rodríguez et al., 2016). The Linguistic Inquiry and Word Count (LIWC) is another dictionary of 6400 words categorized by independent judges into over 70 linguistic dimensions, including standard language categories, psychological processes, relativity-related words and traditional content dimensions (Pennebaker & Graybeal, 2001). AFINN-111 is a dictionary of English words rated for valence with an integer between minus five (negative) and plus five (positive) for 2477 word forms, and can be used for sentiment analysis applications (Nielsen, 2011). The second approach makes use of machine learning techniques and involves building classifiers from labelled instances of texts or sentences through a supervised classification process (Ortigosa et al., 2014). The advantage of machine learning techniques is that they can create trained models for specific contexts. However, they also require the availability of labelled data, which compromises their applicability with new data, as it is the case of shared reviews through eWOMs. Additionally, dictionaries also allow performing a linguistic analysis, which basically measures the percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.

In the context of discourse analysis, a number of studies have been conducted on several well-known online consumer review sites such as Amazon, Tripadvisor (Vásquez, 2012)

or Yelp (Luca, 2011). Amazon, for instance, has attracted a lot of research from marketing, economics and information science. Studies on Amazon reviews from computational fields involve sentiment analysis (Dave et al, 2003) whereas other studies examine online reviews from a linguistic perspective; for example, Pollach (2006) used a methodology which combines elements of case study research, textual analysis and corpus linguistics. It has been argued that pronouns, swear words, negative emotion words, and past tense verbs had correlations with personality profiles (Ziegler et al., 2011). Vásquez (2015) concludes that a closer analysis of discourse is needed in order to develop a more nuanced understanding of this new form of social media. Whereas there is a growing literature on the analysis of consumer opinion websites (Pollach, 2006; Fan & Miao, 2012, etc), not a great deal of research has been conducted on the influence of eWOM from a gender perspective.

The second set of techniques is topic modelling, which consists of extracting the main topics of discussion within a corpus of documents. To this aim, a set of terms are chosen (typically using features such as their TF or TF-IDF values) and then documents are represented as vectors of terms (Paltoglou & Thelwall, 2010). Dimensionality reduction methodologies such as Latent Semantic Indexing (LSI) or Latent Dirichlet Allocation (LDA) are next applied to disambiguate terms and to provide a lower-dimensional representation of documents that better reflects hidden dimensions (Martínez-Torres et al., 2013). These dimensions or topics are finally obtained through unsupervised clustering algorithms. Identification of main themes and topics has been extensively conducted in asynchronous online discussions to obtain the preferences of users. For instance, topic modelling was applied to obtain the preferences of Starbucks consumers

(Martínez-Torres et al., 2015a) and the main topics of discussion in the open innovation community of Dell, called Dell IdeaStorm (Martínez-Torres, 2015). It has also been used to measure the quality of discourse by identifying the main themes and their magnitude in online discussions (Kovanović et al., 2015). Again, very few studies have considered the gender perspective. For instance, Tvinnereim & Fløttum (2015) considered the influence of gender, among other control variables, in answers to open-ended survey questions about climate change.

Finally, the third set of techniques is key words analysis, which consists of extracting relevant terms within a corpus of documents (again using features such as their TF or TF-IDF values) and analyse the relationships among them. As a difference to topic modelling, where the number of topics is unknown a priori, in key word analysis the classes are known in advance. Findings can be graphically visualized using a correspondence analysis, which is a grouping method used for understanding similarities and association between variables (key terms) and classes (Whitlark & Smith, 2001). The resulting graph is the bi-plot, which shows the correspondence between key terms and classes according to their distance to each other, so the closer the items the higher their similarity (Toral et al., 2017). In the context of discourse analysis, key word analysis has been applied to obtain the unique attributes of tourist destinations (Toral et al., 2017) and for authorship identification of online messages (Zheng et al., 2006).

3. Research questions

Content analysis techniques can be used to predict several attributes of authors such as gender, age or even the authorship of a document (Koppel et al., 2004). However, computer-mediated communications are different from other forms of written language typically used in authorship analysis. In the particular case of eWOM communities, it is a one to many communication type, as posted reviews can be read by many people interested in that product or service. Online reviews are more evaluative than narrative as their primary function is to rate, evaluate, describe and provide recommendations to others for or against a particular product or service. Therefore, reviewers tend to show personal preferences and emotions (Kucukyilmaz et al., 2008). Typically, reviewers use the first-person singular when posting reviews, so gender cannot be distinguished by the pronoun “I”. Additionally, and unlike other languages, in English there is no grammatical gender. Although some papers have addressed the topic of authorship identification, they carry out such identification with authors that are regular contributors or follow a specific writing style. For instance, Zheng et al., (2006) considered Internet newsgroup messages and authors posting from 30 to 92 messages, and Abbasi & Chen (2005) studied the specific context of extremist-group web forum messages, using 20 messages for each of 20 authors. However, gender identification in eWOM is a different issue as most of the users only post one or two reviews. As a result, gender prediction must focus on discourse differences regarding preferences and emotions.

Therefore, the first research question can be formulated as follows:

RQ1: Is it possible to distinguish online reviews posted by males and females using a set of frequent key terms?

From the viewpoint of marketing, knowing user preferences is extremely important for customer segmentation. Instead of expensive surveys for market analysis and segmentation, social media is also a useful resource for content analysis, it is less time-consuming and cheaper. Through the Web, it is possible to obtain hundreds or thousands of opinions, so companies and retailers can retrieve information about the strengths and weaknesses of both their products and their competitors. Although there are several online review-based preference measurement models (Decker and Trusov, 2010; Lee and Bradlow, 2011), none of them has considered a gender perspective based on discourse analysis. Gender information can be very useful in recommender systems, which are important and popular tools to analyse users' behaviour over the Web and to generate recommendations regarding their preferences (Mishra et al., 2015). They are typically based on users' past preferences and the sequence of visited web pages, as personal information about users is not usually available. However, registered users in eWOM communities can optionally provide some personal information. Generally, it is compulsory to choose a username and password, but providing data such as gender, age, date of birth, postal address, and country is not. Although many people are reluctant to provide personal data, gender is the exception. As a result, using predictive tools or consulting stored data, gender can be known in advance in most cases, and distinguishing topics of interest by gender can be very useful for recommender systems. In general, gender preferences tends to follow gender stereotypes, and there is a

tendency to maintain separate homogeneous preferences of topics depending on gender (Mohr et al., 2006) or homogeneous decision making (Gorman, 2005). However, Internet anonymity and the gender neutrality of the web postulated in some previous studies can change consumer habits. Through eWOM communities, users can engage in discussions and learn from other users that are hidden behind an alias, so females and males can share comments and opinions about topics traditionally assigned to one of both groups. This leads to the second research question:

RQ2: Which are the main topics of interest for males and females based on the discourse analysis of eWOM content?

4. Methodological approach

Figure 1 details the steps of the methodological approach followed. First, posted reviews by females and males are collected and pre-processed. Then, two dictionary-based text mining techniques and two content analysis techniques are applied to determine the differences of female and male discourse in eWOM communications and to answer the research questions. Next subsections details the main elements of Figure 1.

FIGURE 1

4.1 Data collection and pre-processing

In this study, a specific programme crawler was developed using R, which is a free software environment for statistical computing and graphics (http://www.r-project.org/). The function *readLines()*, which belongs to the base package of R, is used to download the *HTML* data, and the function *htmlParse()* is used to generate an R structure

representing the *HTML* tree. The content of interest is identified and collected using the *XPath* language, which can be easily integrated within *XML* package in R. *XPath* is a language created for doing queries in *XML* content and it is used to find and collect meaningful information with an *XML* document (Gottlob et al., 2003). The function *xpathSApply()*, also belonging to the *XML* package, was used to perform such queries and scrape some specific attributes of the downloaded pages, such as the title and body of the review, the tags that specify the subcategory the posted review belongs to and the user's gender.

Text pre-processing includes removing punctuation and numbers, converting all upper-case to lower-case letters, and a tokenization process to obtain single words. Next, a stop word filter and a stemming process was applied. Stop words are used to remove irrelevant terms and function words (articles, conjunctions, pronouns, etc.). Stemming consists of reducing each word to its basic form (Kayser & Blind, 2017). The stemming framework proposed by Porter (1980) was applied to obtain this common base form.

4.2 Sentiment analysis

In this project, the lexical-based approach is used by means of the Linguistic Inquiry and Word Count (LIWC) dictionary, which is a general-purpose dictionary that can be used to measure various language dimensions, including positive or negative emotions, self-references, causal words, etc. The English version of the LIWC is composed of 6,400 words and word stems (Pennebaker et al., 2015). It distinguishes 21 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.) and many psychological processes that include positive and negative

emotions. Using the tokenized terms after pre-processing the set of collected documents and the dictionary, the number of occurrences per linguistic dimension and psychological process is accounted separately for documents belonging to females and males. Next, a chi-square test is then applied to check if there is a significant difference between females and males regarding that linguistic dimension or psychological process.

4.3 Content analysis

Following the scheme of Figure 1, a bag of words is selected using those tokenized terms with the highest value of TF-IDF metric, which provides a simple model to evaluate the relevance of keywords within a corpus or large collection of documents. TF considers that the importance of a term is proportional to its frequency in the collection of documents, while IDF assumes that the importance of a term relative to a document is inversely proportional to the frequency of occurrence of this term in all the documents (Azam & Yao, 2012). The intuition behind this measure is that the importance of a term to a document is dependent on its frequency as well as the degree of rareness at the document level in the corpus. This condition can also be mixed with some minimum values of frequency and case occurrence. Additionally, an exclusion process may also be applied to remove words that should not be included in the content analysis, such as pronouns, conjunctions, etc., and to remove some words used too frequently or with little discriminative value (i. e. books).

Once a set of words ordered by the TF-IDF value is obtained, a chi-square goodness of fit is applied to check which words are predictors of texts belonging to males or females. Those words exhibiting a statistical significant difference between men and women are

representative of their corresponding discourse, while those other words without statistical difference cannot be associated as belonging to one of them.

Finally, those words exhibiting such statistically difference will be used to build a Naïve Bayes classifier (Mendoza, 2012). This classifier is used to solve a text classification problem, which basically consists of assigning a document *D* to one of a set of *M* pre-defined categories $C=\{c_1,c_2,\ldots,c_M\}$, such as the users' gender. A shared review *d* can be represented as the set of *K* attributes or keywords previously obtained $d=(w_1,w_2,\ldots,w_K)$. The aim of a probabilistic text classifier is to determine the class label *c'* that yields the highest posterior probability *P(c|d)*:

$$c' = \arg\max_{c \in C} \{P(c|d)\}$$

In the case of the Naïve Bayes classifier, *P(c|d)=(P(d)*P(d|c))/P(d).* The computation of *P(d|c)* is difficult due to the huge space of possible documents. However, this computation can be simplified if it is assumed that all the attribute values w_j are independent given the category label *c*. The previous equation can then be reduced to:

$$P(c|d) = P(c) \cdot \frac{\prod_{j=1}^{K} P(w_j|c)}{P(d)}$$

The maximum likelihood Naïve Bayes classifier is obtained by assuming a uniform prior over categories and seeking the optimal category which maximizes the posterior *P(c|d)*:

$$c' = \arg\max_{c \in C} \left\{ P(c) \cdot \prod_{j=1}^{K} P(w_j|c) \right\} = \arg\max_{c \in C} \left\{ \prod_{j=1}^{K} P(w_j|c) \right\}$$

The output metrics to evaluate the classifier will be the precision and recall, which can be computed from the confusion matrix of the binary classification problem.

The obtained bag of words is also used to perform a topic modelling to find hidden topics from a corpus of text (Momeni & Rost, 2016). A frequently used topic modelling algorithm is latent semantic indexing, that applies a singular value decomposition to the original representation of documents as vectors of terms. The dimensionality reduction is achieved after this decomposition by keeping only the largest eigenvalues. The similarity among documents is then measured in this reduced space representation, and a cluster analysis is finally applied to obtain the hidden topics (Thorleuchter & Van den Poel, 2014). An important parameter when performing topic modelling is the selection of the number of topics, which in turns depends on the clustering algorithm. Typically, clustering algorithms select the number of clusters by representing the sum of squared error versus the number of clusters and then applying the elbow criterion. However, when dealing with a corpus of documents, it is also important to apply the analyst judgement (Uys et al., 2008). Some level of granularity is desired but avoiding either superfluous detail or a too coarsely grained solution.

5. Application study

5.1 Data set

The research focuses on Ciao UK, a well-known eWOM website covering a wide variety of products. Ciao is one of the largest eWOM communities in Europe available in local-language versions, with more than 1.3 million members that have posted more than 7 million reviews on 1.4 million of products (Arenas-Márquez et al., 2014, Olmedilla et al., 2016b). The study focuses on the UK site, which has around 44.350 registered users

(Olmedilla et al., 2016a). The users' profile potentially includes personal information about users, but the majority leave most of this information blank. However, gender is perhaps the information most commonly provided. In the case of Ciao UK, over 95% of users provide gender information while for instance only 10% provide information about age. Although the data found in the website is publicly available and can be visualized for free, collecting and structuring all the data requires the use of computational techniques. The huge volume of data makes a manual download unaffordable so a computer programme must be used to perform this task. The collection of data can be automated using a computer programme accessing the source code in HTML. While browsing the website, the computer programme can discriminate the meaningful information and store the relevant variables for the study (Martínez-Torres et al., 2015a).

The data set consists of the title and body of the reviews posted at http://www.ciao.co.uk/ within the category of Books, the tags that specify the subcategory within the generic category of Books, and the gender of the user posting the review. A total of 7450 reviews were collected using the designed crawler. Most of the reviews belong to women, a total of 5367, while only 2083 were posted by men. The larger class (women) was subsampled using random subsampling techniques in order to balance out the dataset (He & Garcia, 2009). The result is a total of 2083 reviews for women and 2083 for men, so both classes are balanced.

The selection of terms (bag of words) was done collecting 200 terms ordered by their TF*IDF value, with the additional conditions of considering only those terms with a frequency higher than 500 and removing those terms occurring in more than 70% of cases. As a result, a list of 200 words with a minimum value of TF*IDF=1504.10 was

obtained. Very frequent and meaningless words, such as ‘book’ or ‘read’ were excluded as they appear in more than 70% of reviews and have no discriminative power between males and females, despite having a TF*IDF value above the threshold. Therefore, considering a maximum frequency is an additional filter.

5.2 Results

The selected dictionary, the LIWC dictionary, classifies English words according to the linguistic dimension they belong to. Therefore, it can be used to differentiate women and men’s discourse in online communications regarding linguistic dimensions. Following the scheme of the upper part of Figure 1, a chi-square test was conducted to study the relationships between the linguistic dimensions and gender. The results are included in Table 1. In those cases where the *p* value is lower than 0.05 (highlighted in grey in Table 1), the null hypothesis can be rejected so an association can be determined between the corresponding linguistic dimension and gender.

TABLE 1

FIGURE 2

To illustrate the association of linguistic dimensions with male and female, a correspondence analysis was conducted (Figure 2). As this study is only considering two categories, male and female, the resulting correspondence analysis is a line (the diagonal of Figure 2) over which the different linguistic dimensions are represented, the left side being male and the right side female. The closer to the left or right side of the diagonal,

the stronger relationship with male and female, respectively. There are also some dimensions that cannot be associated to men or women as the difference is not statistically significant (the ones closer to the centre of the diagonal).

The LIWC dictionary also classifies words attending to the psychological processes they are associated to (second row of Figure 1). Again, a chi-square test was conducted to study the relationships between some psychological processes and gender.

TABLE 2

The statistically significant psychological processes which discriminate between women's and men's discourse are highlighted in grey in Table 2 (*p* value lower than 0.05). Similarly, Figure 3 illustrates the gender association of considered processes. The results show that women are more likely to express positive emotions in their reviews while men tend to be more negative. Women are more likely to choose topics related to friends, sex and family, which are also associated with women preferences regarding books.

FIGURE 3

Our research question asked whether it is possible to distinguish men and women's discourse using a set of key terms. This is the content analysis included in the lower part of Figure 1. The selection of key terms was performed using the TF*IDF criterion, and a key term analysis was conducted following the branch marked with "1" in Figure 1. A chi-square test was used to assess whether the relationships between words and gender are statistically significant, given $p \leq 0.05$. Tables 3 and 4 below show that 127 words out of

the original 200 were selected as statistically significant. Moreover, 81 of them are predictors of female reviews (Table 3) whilst 46 are predictors of male reviews (Table 4).

TABLE 3

TABLE 4

Although the set of words obtained is significantly associated to men and women discourse, the research question *RQ1* asks if it is possible to classify men and women's discourse based on a set of keywords. For this purpose, the whole set of words (127) was used to build a Naïve Bayes classifier in order to determine their predictive power. The results of the Naïve Bayes classifier using a 10 folds cross validation is detailed in Table 5, which shows the confusion matrix and the precision and recall for female, male and total.

TABLE 5

The total precision and recall is around 65%, which means that most of the reviews can be correctly associated to men and women using no other information than a set of key words. Therefore, the online discourse of men and women can be differentiated using a set of keywords. Obviously, the higher the number of keywords, the better the performance of the classifier. However, the main drawback of using a large set of keywords is that it is more difficult to interpret the main topics of preferences of men and women.

The second research question, RQ2, refers to distinguishing the topics of interest for males and females based on the discourse analysis of eWOM content. This is also part

of the content analysis, but using in this case the branch marked with “2” in the lower part of Figure 1. To this end texts associated to women and men were analysed separately by applying the topic modelling technique. Using the bag of words obtained through the TF-IDF value, a singular value decomposition was applied to the documents as vectors of terms in order to measure their similarity in the new space representation. Finally, the k-means clustering algorithm was used to obtain the final number of topics. The selection of the number of topics is based on the elbow criterion but also on the analyst decision. The idea is to keep the coherence and homogeneity of each cluster. As a result, seven and six topics were obtained for females and males, respectively. Tables 6 and 7 show the topics most likely to correspond to females and males respectively alongside some of the keywords for each of the subgroups. Both tables include the name of each cluster, the keywords associated to each of them and the percentage of variance explained by each cluster.

TABLE 6

TABLE 7

The findings show a clear difference in the book topics chosen by women and men. Women are more interested in books about love and character stories, family and children, cookbooks and thrillers, while men prefer historical, humanistic and war books. Women also show an interest in learning through reading while men prefer to obtain some entertainment from reading. The common topic for women and men refers to book recommendations, as expected in online review websites.

To validate the preferences obtained, the posted reviews about books were grouped and counted using the tag system of the books categories within Ciao. Table 8 details the distribution or reviews per tag for females and males.

TABLE 8

The patterns reveal by this table follow the patterns obtained by topic modelling. There is a clear difference towards females regarding children's books, crime and thriller books, romance books, lifestyle books and modern fiction books. Lifestyle books refer to topics like health, beauty, home, food and family, which are clearly distinguished in our analysis as belonging to females' topics. Modern fiction books refer to modern books published by best seller writers. On the contrary, males exhibit preferences for arts & music books (this tag includes films, tv, arts, …), biography books, history books, humour books, science fiction books and sports books. Again, these topics fall within the preferences revealed by topic modelling for men.

6. Discussion and implications

The application of text mining techniques to a large sample collected using a crawler can provide the differences in online discourse and preferences from a gender perspective and it can also be used to compare such differences with those reported for face-to-face communications in previous studies.

One of the findings of this study is that there are differences in the online discourse of men and women and that they show similar differences to those seen in face-to-face communications, as pointed out in previous studies. The prediction that online

communications would be gender neutral grew upon the assumption of anonymity. However, the findings contradict this prediction (Schiffrin et al., 2008), and online discourse tends to replicate gender patterns previously observed in spoken interactions (Zhao et al., 2008). In general, some authors agree that gender discourse must be analysed in the context of the specific online interaction. For instance, women are more active on social networking sites such as Facebook and Twitter while men participate more on music sharing sites (Herring & Stoerger, 2013). In the context of eWOM and books, results from Table 1 shows significant differences in 14 of the 21 linguistic forms analysed. Newman et al (2008) argue that these linguistic functions (nouns, pronouns, verbs, etc) appear to be good markers of how individuals relate to the world. Women appear to use more pronouns than men while men use more swear words (Mehl and Pennebaker, 2003). With regard to the use of auxiliary verbs, the findings of this study appear to correlate with previous research that women use more uncertainty expressions such as seem to, could, may, etc. Women also tend to use more frequently the present tense (Irigaray, 2004). The use of generic mass quantifiers such as some, every, every, all, etc. can also be addressed as part of women discourse while men tend to be more specific with numbers (Coulthard & Coulthard, 2014). Finally, the first person singular is significantly used more by females while the second person appears more frequently in male reviews (Newman et al, 2008; Argamon et al., 2003). However, in contrast to previous studies (Jun, 2012), women use more impersonal pronouns (one, they, it). It is worth mentioning the low rate of use of future tense, which can be explained because reviews refers to past experiences of users, so past and present tenses are more frequent. Additionally, adverbs, conjunctions and function words do not show significant

differences. Function words are made up of pronouns, prepositions, articles, conjunctions and auxiliary verbs (Tausczik and Pennebaker, 2010). It has also been suggested that females and males respond differently when they process and provide information (Richard et al., 2010). As a result, there are also differences in the discourse regarding the different psychological processes, highlighted in this study. As shown in Table 2, most of these processes are associated to women as they appear to talk more about social and affective processes (Mohr et al., 2006).

The present research included a gender classification of online reviews based on a set of attributes or keywords. Attributes can be user-specific or message-specific (Kucukyilmaz et al., 2008). As we assume no previous knowledge about users, who use an alias, attributes are message-specific and they were selected attending to their TF*IDF values. Findings reveal that it is possible to classify accurately documents belonging to men and women using a small set of attributes which answers our first research question. The possibility of distinguishing online reviews posted by males and females using a set of frequent key terms means that there is no gender neutrality on the web, despite authors being hidden behind an alias.

The association of keywords and gender also reveals gender preferences related to books. A content analysis was carried out to extract the main topics of interest for women and men in the category of books. Results revealed different preferences regarding thematic approach and aim of the reading, more learning-oriented in the case of women and more entertainment-oriented in the case of men. These results tend to follow gender stereotypes also present in the offline world (Mohr et al., 2006). From a practical

perspective, Figure 4 summarizes those words that can help marketers when targeting females and males.

As a difference to previous studies related to gender discourse, this study considers a variety of text mining techniques to highlight the differences between males and females regarding linguistic forms, psychological processes and content. The three analyses concluded with significant differences between males and females, despite the anonymity provided by the Internet.

In general, gender is one of the most commons forms of segmentation used by marketers and advertisers. Previous studies have concluded that males and females differ in information process and decision making (Kim et al., 2007). Therefore, understanding online users' preferences from a gender perspective has important implications for online retailers (Hou & Elliott, 2016). For instance, marketers may benefit by creating gender-sensitive website content and presentation, and advertisers could develop gender-sensitive online communication strategies. Both the content and the presentation of the message could incorporate some of the findings of this study, like addressing the thematic preferences of males and females, or even using a linguistic style closer to the one they like when making reviews.

Another important implication of this study is the possibility of predicting the author's gender based on a set of keywords retrieved from the specific context where discussions took place. This is quite important for online recommender systems, which today represent a key piece in online shopping due to their ability to deliver shopping advice, stimulate consumers' purchase desires and boost sales (Hung, 2005). Previous studies point out that gender plays a moderation role on consumers' acceptance of recommender

system (Doong and Wang, 2011; Martínez-Torres et al., 2015b). Therefore, knowing in advance the author's gender can help recommender systems to provide more accurate advice. Whenever this information is not available, it can be predicted using the methodology proposed in the study.

6.1 Limitations

This study is limited to a specific website, Ciao UK, and a specific category of products, Books. However, the data collection and analysis method could be extended to other product categories as ciao.com has 28 different product categories, so the present findings could be compared with the rest of categories. Moreover, products can be generally classified as either search products or experience products (Cui et al., 2012). Search products are those products that can be evaluated through objective standards while experience products are typically evaluated by affective attributes. The specific relationship between gender and product category or typology could be further analysed.

Another limitation of this work revolves around considering key terms as individual words. The main disadvantage is polysemy, that is, the diversity of meanings a word can have. An alternative consists of using phrases (group of words) that reduces the effect of polysemy, but at the cost of diminishing the importance of some terms when used in different contexts or associated to other terms (Dascalu, 2014).

The number of selected key words is also a parameter that can be further studied. In general, this parameter must balance the accuracy of the identification and the interpretability of results when conducting the topic analysis (Monay & Gatica-Perez,

2003). A parametric analysis considering different values of the number of keywords can help to select an optimum value.

Finally, the linguistic dimension and psychological processes were analysed using a dictionary. As a different approach, machine-learning techniques working on a set of previously annotated reviews could be used to obtain a context specific dictionary (Ortigosa et al., 2014).

7. Conclusions

This paper contributes to the debate about gender neutrality of the web. The application of a set of text mining techniques over the collected online reviews belonging to the category of Books within eWOM Ciao UK reveals significant differences between males and females in the linguistic domain, the psychological processes and content. Therefore, it can be concluded that the Internet anonymity is not facilitating the gender neutrality of the web, as claimed by some authors. However, these differences can be exploited by online retailers to provide personalized recommendations or by marketers and advertisers to improve the communication with customers considering gender as another input variable. The possibility of predicting gender, as demonstrated by this study, also opens the way to more customized recommender systems.

Acknowledgements

This work was supported by the Consejería de Economía, Innovación, Ciencia y Empleo under the Research Project with reference P12-SEJ-328 and by the Programa

Estatal de Investigación, Desarrollo e Inovación Orientada a los Retos de la Sociedad under the Research Project with reference ECO2013-43856-R.

References

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 20(5), 67-75.

Arenas-Márquez, F. J., Martínez-Torres, M. R., Toral, S. L. (2014). Electronic word-of-mouth communities from the perspective of social network analysis, Technology Analysis & Strategic Management, 26 (8): 927-942.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts, Interdisciplinary Journal for the Study of Discourse. 23(3), 321-346.

Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization, Expert Systems with Applications, 39(5), 4760-4768.

Blank, T. O., & Adams-Blodnieks, M. (2007). The who and the what of usage of two cancer online communities. Computers in human behavior, 23(3), 1249-1257.

Brownlow, S., Rosamond, J. A., Parker, J. A. (2003). Gender-linked linguistic behavior in television interviews, Sex Roles, 49 (3-4), pp. 121-132.

Bucholtz, M. (2003) Theories of discourse as theories of gender: Discourse analysis in language and gender studies. In The handbook of language and gender, J. Holmes and M. Meyerhoff, (Eds). Oxford Blackwell, pp. 43-68.

Burri, M., Baujard, V., & Etter, J. F. (2006). A qualitative analysis of an internet discussion forum for recent ex-smokers. Nicotine & Tobacco Research, 8(Suppl_1), S13-S19.

Colley, A., Todd, Z., Bland, M., Holmes, M., Khanom, M. & Pike, H. (2004). Style and content in e-mails and letters to male and female friends, Journal of Language and Social Psychology, 23 (3) 369-378.

Coulthard, M., & Coulthard, M. (2014). An introduction to discourse analysis. Routledge.

Cui, G., Lui, H. K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. International Journal of Electronic Commerce, 17(1), 39-58.

Dascalu, M. (2014). Computational Discourse Analysis. In Analyzing Discourse and Text Complexity for Learning and Collaborating (pp. 53-77). Springer International Publishing.

Dave, K. et al (2003) Mining the Peanut Gallery: Opinion extraction and semantic classification of product reviews, Proceedings of the 12th International Conference on World Wide Web, New York

Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews, International Journal of Research in Marketing, 27(4), 293-307.

Doong, H. S., Wang, H. C. (2011). Do males and females differ in how they perceive and elaborate on agent-based recommendations in Internet-based selling?, Electronic Commerce Research and Applications, 10, 595–604.

Dumay, J., & Cai, L. (2014). A review and critique of content analysis as a methodology for inquiring into IC disclosure. Journal of intellectual capital, 15(2), 264-290.

Dwight, J. (2004). 'I'm Just Shy': Using Structured Computer-Mediated Communication to Disrupt Masculine Discursive Norms. E-Learning and Digital Media, 1(1), 94-104.

Fan, Y.-W., Miao, Y.-F. (2012). Effect of electronic word-of-mouth on consumer purchase intention: The perspective of gender differences, International Journal of Electronic Business Management, 10 (3) 175-181.

González-Rodríguez, M. R., M. R. Martínez Torres, and S. Toral. (2016). Post-visit and pre-visit tourist destination image through eWOM sentiment analysis and perceived helpfulness. International Journal of Contemporary Hospitality Management, 28(11): 2609-2627.

Gorman, E. H. (2005). Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms. American Sociological Review, 70(4), 702-728.

Gottlob, G., Koch, C., & Pichler, R. (2003). The complexity of XPath query evaluation. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 179-190). ACM.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data, IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

Hennig-Thurau, T. et al (2004) “Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?” Journal of Interactive marketing, vol 18, pp.38-52

Herring, S. C. (2003). Gender and power in on-line communication in The Handbook of language and gender, 202-228.

Herring, S. C., & Stoerger, S. (2013). Gender and (a)nonymity in Computer-Mediated Communication in Erlich et al (eds.) The Handbook of Language, Gender, and Sexuality 2nd edition, 567-586.

Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U., & de Jong, F. (2014). Multi-lingual support for lexicon-based sentiment analysis guided by semantics, Decision support systems, 62, 43-53.

Hollenbaugh, E., and. Everett, M. (2013) The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect, Journal of Computer-Mediated Communication 18 (3), 283-302.

Hou, J., & Elliott, K. (2016). Gender differences in online auctions. Electronic Commerce Research and Applications, 17, 123-133.

Hung, L. P. (2005). A personalized recommendation system based on product taxonomy for one-to-one marketing online. Expert Systems with Applications, 29, 2, 383–392.

Irigaray, L. (Ed.). (2004). Luce Irigaray: key writings. A&C Black.

Jun, Z. Y. (2012). Gender identity and language usages in masculine-feminine discourse, GSTF Journal of Law and Social Sciences (JLSS), 2(1), 201.

Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. Technological Forecasting and Social Change, 116, 208-215.

Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme, Decision Support Systems, 57, 245-257.

Kim, D. Y., Lehto, X. Y., & Morrison, A. M. (2007). Gender differences in online travel information search: Implications for marketing communications on the internet. Tourism management, 28(2), 423-433.

Kim, K., Park, O. J., Yun, S., & Yun, H. (2017). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart

destination management. Technological Forecasting and Social Change, doi 10.1016/j.techfore.2017.01.001

Koppel, M., Argamon, S., & Shimoni, A. R. (2004). Automatically categorizing written texts by author gende,. Computing Reviews, 45(1), 43.

Kovanović, V., Joksimović, S., Gašević, D., Hatala, M., & Siemens, G. (2015). Content Analytics: the definition, scope, and an overview of published research. Handbook of Learning Analyitcs, 77-92.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication, Information Processing & Management, 44(4), 1448-1466.

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. Journal of Marketing Research, 48(5), 881-894.

Luca, M. (2016) User-generated content and social media in Anderson, S et al (eds.) Handbook of media Economics, North-Holland Publishing

Mackiewicz, J. (2010) “Assertions of Expertise in Online Product Reviews.” The Journal of Business and Technical Communications,

Martínez-Torres, M. R., Toral, S. L., Barrero, F., Gregor, D. (2013). A text categorisation tool for open source communities based on semantic analysis, Behaviour & Information Technology, 32 (6) 532-544.

Martínez-Torres, M. R. (2014). Analysis of open innovation communities from the perspective of Social Network Analysis. Technology Analysis & Strategic Management, 26(4), 435-451.

Martínez-Torres, M. R., Diaz-Fernandez, M. C. (2014). Current issues and research trends on open-source software communities, Technology Analysis & Strategic Management, 26 (1) 55-68.

Martínez-Torres, M. D. R., Rodriguez-Piñero, F., & Toral, S. L. (2015a). Customer preferences versus managerial decision-making in open innovation communities: the case of Starbucks. Technology Analysis & Strategic Management, 27(10), 1226-1238.

Martínez-Torres, M. R. (2015). Content analysis of open innovation communities using latent semantic indexing, Technology Analysis & Strategic Management, 27 (7) 859-875.

Martínez-Torres, M. R., Díaz-Fernández, M. D. C., Toral, S. L., & Barrero, F. (2015b). The moderating role of prior experience in technological acceptance models for ubiquitous computing services in urban environments. Technological Forecasting and Social Change, 91, 146-160.

Martinez-Torres, R., & Olmedilla, M. (2016). Identification of innovation solvers in open innovation communities using swarm intelligence. Technological Forecasting and Social Change, 109, 15-24.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. Journal of personality and social psychology, 84(4), 857.

Mendoza, M. (2012). A new term-weighting scheme for naïve Bayes text categorization. International Journal of Web Information Systems, 8(1), 55-72.

Mishra, R., Kumar, P., & Bhasker, B. (2015). A web recommendation system considering sequential information. Decision Support Systems, 75, 1-10.

Momeni, A., & Rost, K. (2016). Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. Technological Forecasting and Social Change, 104, 16-29.

Mo, P. K., Malik, S. H., & Coulson, N. S. (2009). Gender differences in computer-mediated communication: a systematic literature review of online health-related support groups. Patient education and counseling, 75(1), 16-24.

Monay, F., & Gatica-Perez, D. (2003). On image auto-annotation with latent space models. In Proceedings of the eleventh ACM international conference on Multimedia (pp. 275-278). ACM.

Mohr, K. A. (2006). Children's choices for recreational reading: A three-part investigation of selection preferences, rationales, and processes. Journal of Literacy Research, 38(1), 81-104.

Moss, G., Hamilton, C., & Neave, N. (2007). Evolutionary factors in design preferences. Journal of Brand Management, 14(4), 313-323.

Mulac, A., & Lundell, T. L. (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. Language & Communication, 14(3), 299-309.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211-236.

Nielsen, F. A. (2011). A new ANEW: Evaluation of word list for sentiment analysis in microblogs, In Proc. of the ESW2011 Workshop on 'Making sense of Microposts: Big

Things come in small packages, no. 78 in CEUR Workshop Proceedings, Heraklion, pp. 93-98.

Olmedilla, M., Martínez-Torres, M. R., & Toral, S. L. (2016a). Harvesting Big Data in social science: A methodological approach for collecting online user-generated content. Computer Standards & Interfaces, 46, 79-87.

Olmedilla, M., Martinez-Torres, M. R., Toral, S. (2016b). Examining the Power Law Distribution among eWOM communities: A characterization approach of the Long Tail, Technology Analysis & Strategic Management, 28 (5), 601-613.

Ortigosa, A., Martín, J. M., Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning, Computers in Human Behavior, 31, 527-541.

Paltoglou, G., & Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1386-1395). Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Park, J., Yoon, Y., & Lee, B. (2009). The effect of gender and product categories on consumer online information search. ACR North American Advances.

Pennebaker, J. W. & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration, Current Directions in Psychological Science, 10 (3) 90-93.

Pennebaker, J. W., Mehl, M. R., Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves, Annual Review of Psychology, 54 (1) 547-577.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015, Repository Home, University of Texas, http://hdl.handle.net/2152/31333

Pollach, I. (2006) Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites, Proceedings of the 39th Hawaii International Conference on System Sciences.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Richard, M.-O., Chebat, J.-C., Yang, Z., & Putrevu, S.(2010).A proposed model of online consumer behavior: Assessing the role of gender. Journal of Business Research, 63, 926–934

Schiffrin, D., Tannen, D., & Hamilton, H. E. (Eds.). (2008). The handbook of discourse analysis. John Wiley & Sons.

Sproull, L., Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication, Management science, 32 (11) 1492-1512.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. British Journal of Social Psychology, 40, 193–208.

Thorleuchter, D., & Van den Poel, D. (2014). Quantitative cross impact analysis with latent semantic indexing. Expert Systems with Applications, 41(2), 406-411.

Toral, S. L., Martínez-Torres, M. R., & Gonzalez-Rodriguez, M. R. (2017). Identification of the Unique Attributes of Tourist Destinations from Online Reviews. Journal of Travel Research, doi 10.1177/0047287517724918.

Tvinnereim, E., & Fløttum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. Nature Climate Change, 5(8), 744.

Ulbrich, F., Christensen, T., Stankus, L. (2011). Gender-specific on-line shopping preferences, Electronic Commerce Research, 11 (2) 181-199.

Uys, J. W., Du Preez, N. D., & Uys, E. W. (2008, July). Leveraging unstructured information using topic modelling. In IEEE International Conference on Management of Engineering & Technology, 2008. PICMET 2008. Portland, pp. 955-961.

Vásquez, C. (2012). Narrativity and involvement in online consumer reviews: The case of TripAdvisor, Narrative Inquiry 22 (1), 105-121

Vásquez, C. (2015), The discourse of online consumer reviews, Bloomsbury

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction. A relational perspective. Communication research, 19 (1) 52-90.

Wang, Y., & Rodgers, S. (2010). Electronic word of mouth and consumer generated content: From concept to application. Handbook of research on digital media and advertising: user generated content consumption, edited by Matthew S. Eastin, et al, 212-231.

Whitlark, D. B., & Smith, S. M. (2001). Using correspondence analysis to map relationships. Marketing Research, 13(3), 22.

Xia, X. (2013). Gender Differences in Using Language, Theory and Practice in Language Studies, 3 (8) 1485-1489.

Xiao, S., Wei, C. P., & Dong, M. (2016). Crowd intelligence: Analyzing online product reviews for preference measurement. Information & Management, 53(2), 169-182.

Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. Information & Management, 52(7), 850-858.

Zhang, Y., Dang, Y., Chen, H. (2013). Examining gender emotional differences in Web forum communication, Decision Support Systems, 55 (3) 851-860.

Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on Facebook: digital empowerment in anchored relationships. Computers in Human Behavior, 24(5), 1816-1836.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the Association for Information Science and Technology, 57(3), 378-393.

Ziegler, M., MacCann, C., & Roberts, R. (2011). New perspectives on faking in personality assessment. Oxford University Press.

Tables:

	Female	*Male*	*Chi-square test*		*Predict*
			Chi sq.	*p*	
Total Pronouns	2771	2248	54.248	0.000	Female
Swear Words	505	637	15.321	0.000	Male
Quantifiers	19145	18229	22.013	0.000	Female
Present Tense	18104	13231	755.478	0.000	Female
Past Tense	640	462	28.666	0.000	Female
Auxiliary Verbs	166	74	35.223	0.000	Female
Common Verbs	18624	13483	820.716	0.000	Female
Impersonal Pronouns	2692	2165	56.929	0.000	Female
Negations	213	139	15.521	0.000	Female
Articles	223	302	11.926	0.001	Male
1st person Singular	37	14	10.362	0.001	Female
Prepositions	503	408	9.861	0.002	Female
Numbers	2429	2644	9.215	0.002	Male
2nd person singular	29	53	7.036	0.008	Male
Adverbs	2761	2907	3.831	0.050	Male
Function Words	19453	19088	3.284	0.070	Female
Future Tense	52	43	0.848	0.357	Female
3rd Singular	12	15	0.335	0.563	Male
Personal Pronouns	79	83	0.101	0.751	Male
Conjunctions	47	46	0.010	0.919	Female
3rd Plural	1	1	0.000	1.000	Female

Table 1. Linguistic dimension of shared opinions by gender.

	Female	*Male*	*Chi-square test*		*Predict*
			Chi sq.	*p*	
Family	10298	5647	1354.420	0.000	Female
Sexual	4951	2983	487.211	0.000	Female
Positive Emotion	50505	44508	375.644	0.000	Female
Anger	9204	11590	274.927	0.000	Male
Friends	3376	2302	202.633	0.000	Female
Affective Processes	78378	73636	145.657	0.000	Female
Relativity	85125	89007	88.416	0.000	Male
Body	7221	6521	35.322	0.000	Female
Negative Emotion	27496	28843	32.855	0.000	Male
Anxiety	5736	5244	21.810	0.000	Female
Time	40789	41770	12.132	0.000	Male
Certainty	8817	9246	10.396	0.001	Male
Health	9085	8702	8.064	0.005	Female
Sadness	5481	5223	6.095	0.014	Female
Discrepancy	3435	3234	5.962	0.015	Female

Table 2. Psychological processes of shared opinions by gender.

Female words	*Chi-square test*		*Female words*	*Chi-square test*		*Female words*	*Chi-square test*	
	Chi sq	*p*		*Chi sq*	*p*		*Chi sq*	*p*
Love	219.13	0.0000	Character	53.64	0.0000	Kind	16.87	0.0000
Price	178.08	0.0000	Amazon	51.73	0.0000	Dog	16.19	0.0001
Children	148.15	0.0000	Friend	51.43	0.0000	Word	15.66	0.0001
Child	141.49	0.0000	Feel	48.48	0.0000	Author	14.41	0.0001
Girl	131.43	0.0000	Son	47.32	0.0000	Favourite	13.79	0.0002
Husband	126.42	0.0000	Recipe	47.05	0.0000	Clear	12.69	0.0004
Cover	126.33	0.0000	Lot	46.99	0.0000	Local	12.40	0.0004
Recommend	99.96	0.0000	Friend	46.61	0.0000	Person	11.55	0.0007
Easy	97.41	0.0000	Family	45.34	0.0000	Life	11.40	0.0007
Mother	93.75	0.0000	Sister	42.84	0.0000	Information	10.77	0.0010
ISBN	93.31	0.0000	Happy	42.21	0.0000	Plot	9.25	0.0023
Paperback	83.09	0.0000	Young	38.83	0.0000	Reader	8.96	0.0028
Reading	74.69	0.0000	School	36.21	0.0000	Follow	8.72	0.0031
Bit	74.06	0.0000	Page	27.23	0.0000	Whilst	8.65	0.0033
Felt	73.69	0.0000	Thought	25.49	0.0000	Start	8.33	0.0039
Parent	72.34	0.0000	House	25.04	0.0000	Live	8.25	0.0041
Woman	70.88	0.0000	Buy	24.77	0.0000	Character	7.91	0.0049
Daughter	70.48	0.0000	Simple	24.66	0.0000	Job	7.86	0.0051
Baby	69.85	0.0000	Relationship	24.51	0.0000	Text	6.55	0.0105
Thing	67.06	0.0000	Women	23.68	0.0000	Language	6.40	0.0114
Live	61.82	0.0000	Fun	23.13	0.0000	Night	6.13	0.0133
Understand	60.14	0.0000	Food	21.53	0.0000	Crime	5.66	0.0173
Enjoy	60.10	0.0000	Ending	21.27	0.0000	Chapter	5.48	0.0192
Story	58.38	0.0000	Year	19.58	0.0000	Thriller	5.48	0.0192
Copy	57.12	0.0000	Nice	18.67	0.0000	Worth	5.42	0.0200
Learn	54.52	0.0000	Difficult	18.01	0.0000	Dark	4.93	0.0264
Home	53.92	0.0000	Picture	17.51	0.0000	UK	4.69	0.0304

Table 3. Words predicting female gender

Male words	*Chi-square test*		*Male words*	*Chi-square test*	
	Chi sq	*p*		*Chi sq*	*p*
War	181.99	0.0000	Men	17.57	0.0000
British	163.72	0.0000	Death	16.03	0.0001
John	114.65	0.0000	Reader	13.74	0.0002
King	82.43	0.0000	London	12.82	0.0003
History	77.00	0.0000	Day	12.76	0.0004
Early	73.49	0.0000	Film	11.79	0.0006
World	63.40	0.0000	Detail	10.27	0.0014
Number	58.26	0.0000	Previous	9.79	0.0018
Music	57.91	0.0000	Write	8.86	0.0029
American	50.73	0.0000	Mind	8.46	0.0036
TV	46.11	0.0000	Series	8.35	0.0039
Man	44.95	0.0000	Work	7.78	0.0053
Order	43.71	0.0000	City	7.66	0.0056
Human	36.97	0.0000	Black	7.50	0.0062
James	30.78	0.0000	Style	6.83	0.0090
Fact	26.00	0.0000	Time	6.49	0.0109
Writer	25.54	0.0000	Great	6.38	0.0116
Event	23.12	0.0000	Age	5.44	0.0197
Society	23.05	0.0000	Section	4.96	0.0260
Case	22.90	0.0000	Place	4.94	0.0263
Country	21.37	0.0000	Interesting	4.43	0.0354
Year	18.98	0.0000	Left	4.17	0.0411
Short	18.05	0.0000	Interest	4.10	0.0430

Table 4. Words predicting male gender.

	Predicted		*Total*	*Precision*	*Recall*
	Female	*Male*			
Female	1532	552	2084	0.6307	0.7351
Male	897	1187	2084	0.6824	0.5694
Total	2429	1739	7450	0.6566	0.6522

Table 5. Naïve Bayes results.

	Topics	*Keywords*	*%Var*
1	Love & character stories	Character; Story; Feel; Lot; Plot; Enjoy; Bit; Start; Life; Author; Recommend; Love; Friend; Year; Page	9.25
2	Recommendations	ISBN; Paperback; Price; UK; Amazon; Cover; Copy; Page; Author	3.07
3	Learning	Language; Understand; Clear; Learn; Difficult; Text; Kind	6.31
4	Family	Mother; Family; Daughter; Husband; Sister; Woman; Child; Life; Home; Live; Young; Year; House; Baby; Girl	5.26
5	Cookbooks	Recipe; Food; Buy; Information; Easy; Follow	2.99
6	Thriller	Thriller; Crime; Plot	2.59
7	Infant	Fun; Picture; Child; Favourite; Simple; Children; Dog; Text; Baby	2.04

Table 6. Clusters representing the main distinctive topics of interest for females.

	Name	*Keywords*	*%Var*
1	Recommendations	Time; Write; Interest; Fact; Reader; Great; Detail; Style; Section; Number; Place; Case; Short; Mind; Series	14.77
2	War	War; Country; British; History; World; Society; American	5.09
3	Humanistic	Human; Death	4.20
4	Historical	King; James; Previous; Series	2.84
6	Entertainment	Film; Music; Tv; John; London	2.39

Table 7. Clusters representing the main distinctive topics of interest for males.

Tag	Female	Male	**Tag**	Female	Male

Arts & Music Books	83	22	Humour Books	87	19
Biography Books	263	112	Information Technology Books	23	3
Business & Finance Books	24	5	Language & Linguistics Books	13	12
Psychology/Sociology Books	13	8	Lifestyle Books	58	138
Political Books	20	3	Comics	15	2
Reference Books	25	22	Modern Fiction Books	288	540
Children's Books	175	424	Philosophy Books	13	1
Classics	41	24	Religious	12	4
Crime Books	125	181	Romance Books	7	66
Fantasy Books	74	99	Science Fiction Books	107	22
Gay & Lesbian Books	7	3	Sport Books	67	0
Graphic Novels	20	5	Thriller Books	107	170
History Books	225	66	Travel Books	61	41
Horror Books	51	26	Others	79	65

Table 8. Distribution of reviews using the tag system of books categories at Ciao UK.

Figures:

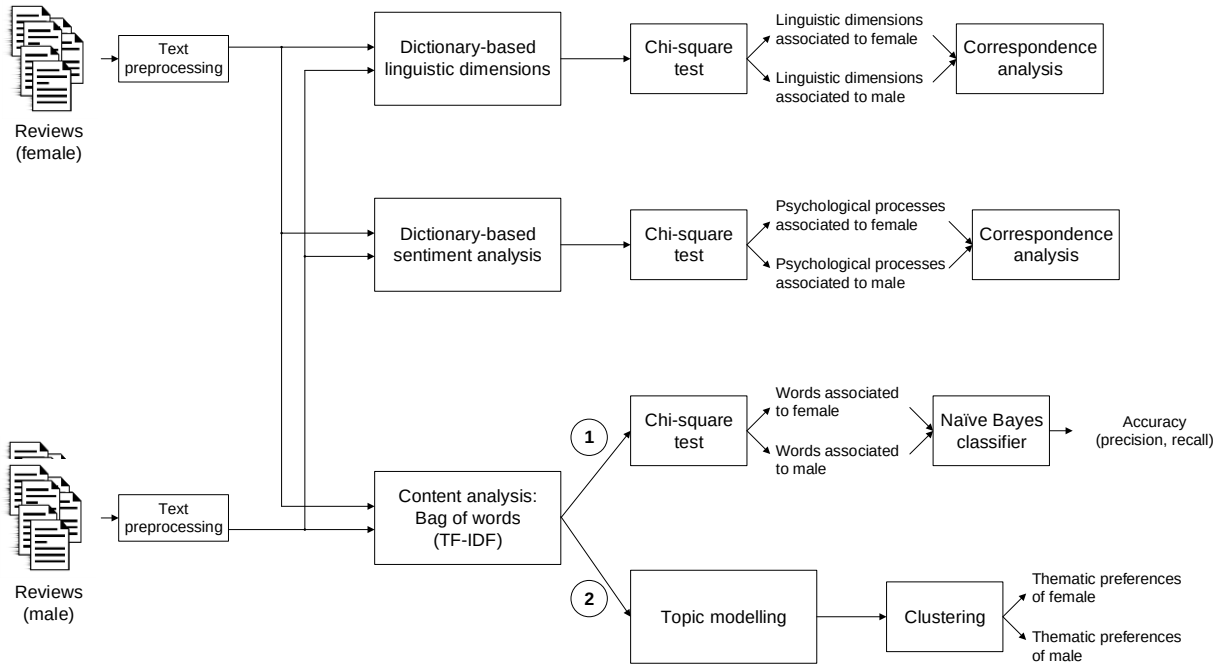

Figure 1. Block diagram of the methodological approach.

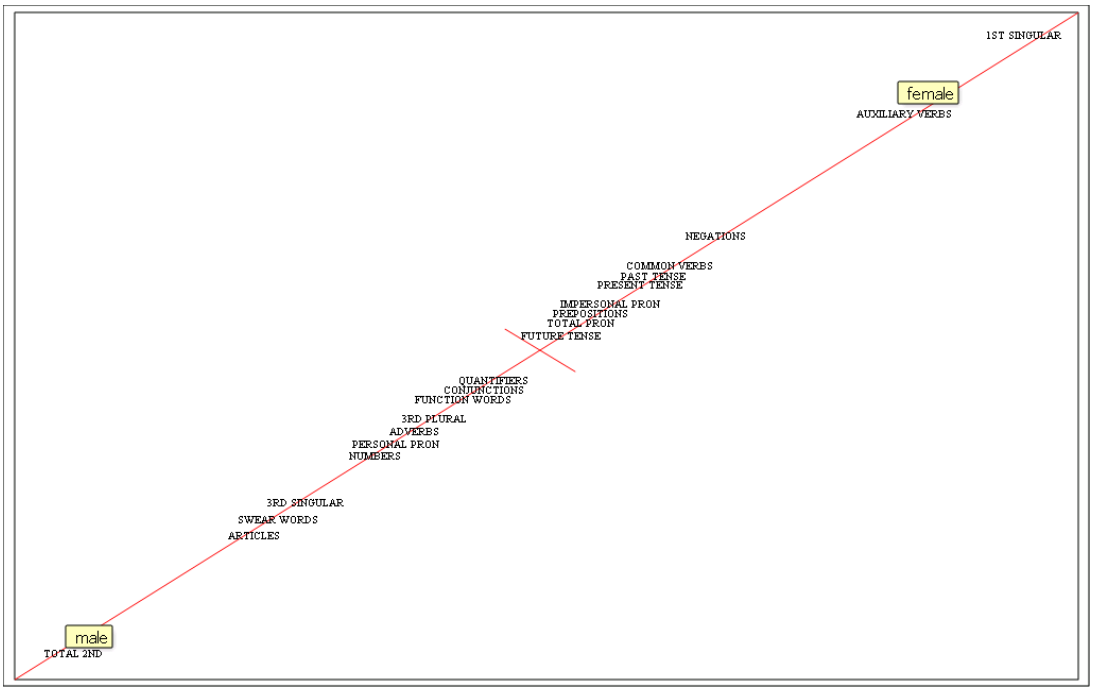

Figure 2. Distribution of the linguistic dimensions over a bi-dimensional plane.

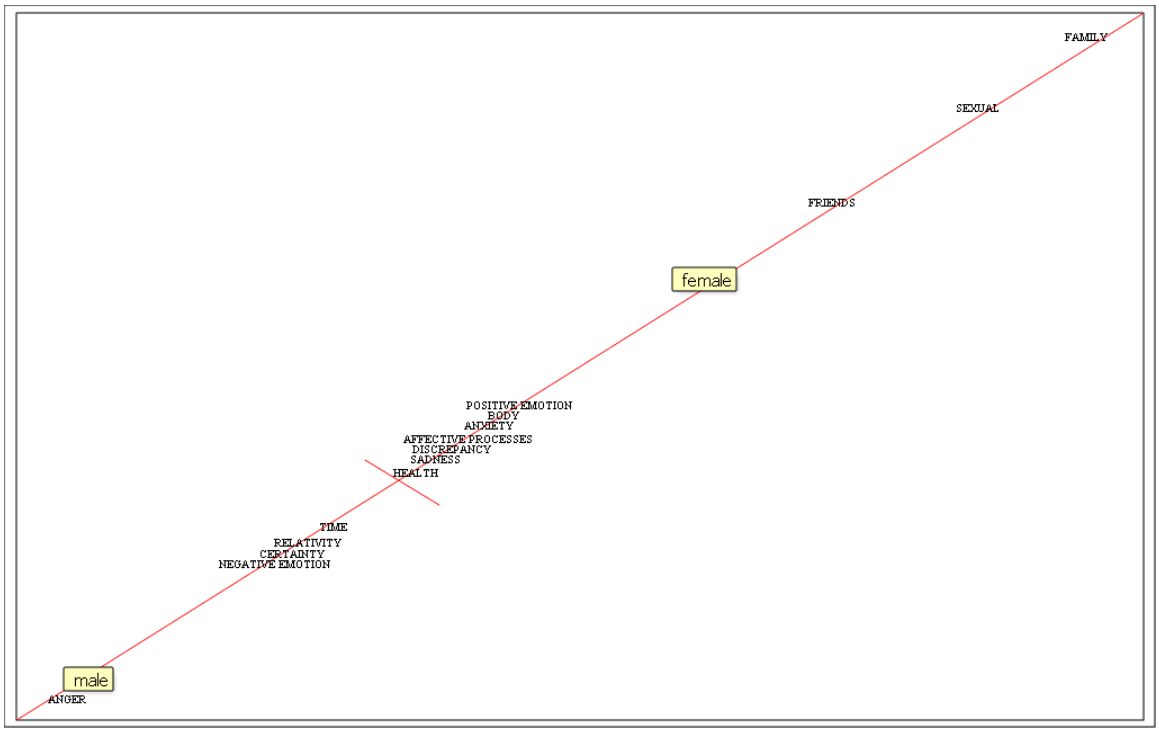

Figure 3. Distribution of the psychological dimensions over a bi-dimensional plane.

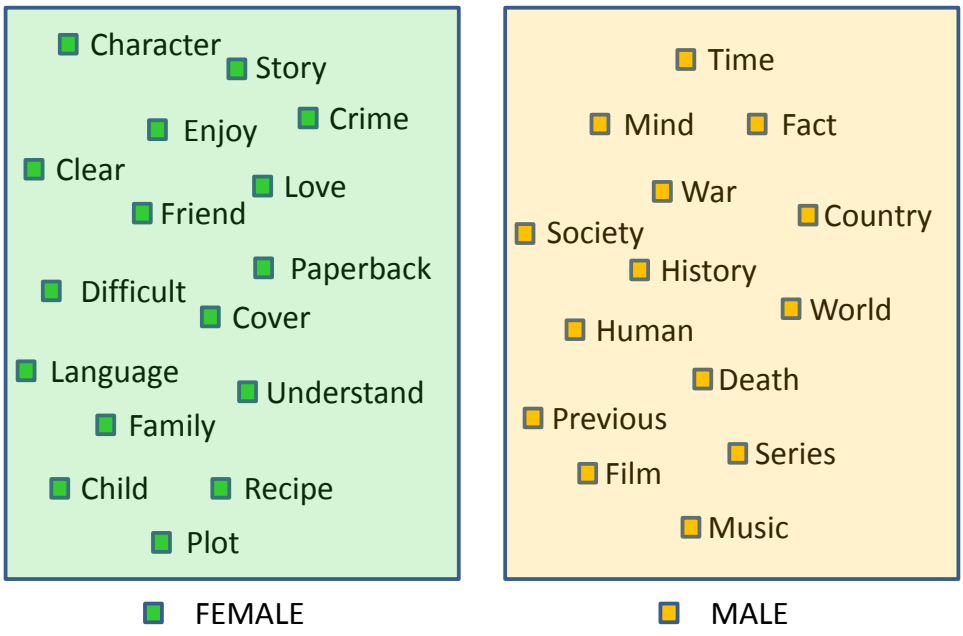

Figure 4. Summary of words that can help marketers when targeting females and males.