A text mining and topic modelling perspective of ethnic marketing research

Sérgio Moro[1,2,*], Guilherme Pires[3], Paulo Rita[4] and Paulo Cortez[2]

Abstract

This study presents an enhanced automated approach based on literature analysis and synthesis for establishing the dimensions of the ethnic marketing literature, covering a set of 239 journal articles published by nine major publishers. The approach reported is enhanced by two novel procedures to address previously identified limitations, namely: definition of a relevant dictionary based on both a sufficient lexicon extracted from a definition of the core theme and a conditional dictionary, with related but non-core terms; and a visually appealing pictorial representation to summarize the discovered topics.

The application of the method to ethnic marketing indicates that ethnic marketing research is characterised by high conceptual heterogeneity, although a clear definition of "ethnic marketing" is imperative for research development. Overall, the paper advances an approach with considerable scalability advantages when compared with extant approaches, an important issue to consider when textual sources become big data.

Keywords

Ethnic marketing; literature analysis and synthesis; literature dimensions; filtering dictionaries; text mining; topic modelling.

---

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal
[2] ALGORITMI Research Centre, University of Minho, Portugal
[3] Newcastle Business School, Faculty of Business & Law, University of Newcastle, Australia
[*] Corresponding author. Email: scmoro@gmail.com; Address: Av. Forças Armadas, 1649-026 Lisboa, Portugal
[4] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal & NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal.

1. Introduction

The increasing importance afforded to ethnicity and to minority ethnic groups, referred to as the new majority in the US (Armstrong, 2013), reflects not only the potential marketing opportunities offered by ethnic marketing to businesses in both advanced and developing economies, but also an expectation that business will take effective advantage of those opportunities. Such opportunities are widely acknowledged in the specialist literature (Jamal et al., 2015; Pires & Stanton, 2015) and expected to expand given increasing globalization trends and growing ethnic diversity of countries around the globe (Cleveland et al., 2015). Yet, ethnic marketing research, and its conversion into teaching and practice, remains largely a niche area of interest, a curiosity open to conjecture. The question is why is that so?

Since the importance of ethnic consumers and minority ethnic groups for marketing purposes is not questioned in the literature, the position subscribed in this paper is that the rigidities affecting the development of ethnic marketing research may need to be understood and demonstrated, convincingly and unambiguously, before they can be overcome. The challenge in addressing the rigidities is to devise a method of analysis that is perceived to provide independent and reliable outcomes, at arm's length of the analysts' preconceptions or ideologies. Focusing on purposely set criteria to search and select published research, this study combines the application of the text mining technique with topic modelling procedures to provide a snapshot of the research that was conducted on ethnic marketing in the period from 2005 to 2015. Contrasting from critical reviews of the literature that are open to confirmatory bias (Basuroy et al., 2003), using a literature analysis allows for unbiased conclusions to be derived, potentially highlighting the manifestation of the rigidities in ethnic marketing research development and facilitating remedial actions.

Data mining has been successfully applied as a valuable tool in predicting the impact of social media on brand building (Moro et al., 2016), and in modelling online feedback score of hotels (Moro et al., 2017). But it is a tool that requires structured data. Text mining overcomes this limitation because it allows for unstructured data, such as text documents, to be parsed. Hence, in order to facilitate effective literature analysis and summarization, text mining and topic modelling have recently been applied in marketing, management and information technology contexts, namely cause-related marketing (Guerreiro et al., 2016), banking (Moro et al., 2015), and cloud computing (Upreti et al., 2016). The procedure has no precedent when applied to ethnic marketing literature.

Prompted by the considerations about ethnic marketing outlined above, preliminary examination of the ethnic marketing literature by the existing text mining tool revealed a degree of complexity that indicated a need to improve on existing methods to select the literature to be analysed. This is a major objective for the current research.

The main contributions of this paper are as follows:

• performing a semi-automated literature analysis and synthesis based on text mining and topic modelling on ethnic marketing, concomitantly improving the literature analysis method through a clearer definition of domain knowledge dictionaries;

• uncovering the topic model built through a novel visual concept map that allows exhibition of the key aspects on ethnic marketing, providing a clear picture on the inherent main trends and relationships;

• discussing the unveiled topics based on the frequency of recurring terms as revealed by specifically designed word clouds;

- evaluating the proposed hypotheses on the lack of a clear understanding of the dimensions of the "ethnic marketing" discipline within the academic literature, including recommendations on how to address this issue.

2. Theoretical background

2.1. Ethnic marketing research

Research on a particular topic develops over time through incremental learning (Eisenhardt, 1989), by drawing on research findings as a stepping-stone for new research. A clear topic definition is usually required, and any confounding interpretations are undesirable. Such is the case, for example, of ethnic marketing research that does not distinguish between ancestry, ethnicity, ethnic identity, culture and cultural diversity. It is also the case when cross-cultural marketing and/or multicultural marketing are used interchangeably with ethnic marketing. Although it may be argued that all those terms have some commonality with ethnicity (ABS, 2011), different meanings and implications make it imperative to distinguish between those terms.

Broadly identified as 'differentiated marketing towards an ethnic group' (Cui, 1998:88), ethnic marketing is concerned with individual ethnic groups and their institutions, whether the majority or the minority, within countries and across borders if relevant (e.g. the Portuguese in Australia and in Canada, or the Iranian in Australia and in the UK). It is well known that definitional differences (such as those verifiable for 'ethnicity' and for 'ethnic marketing') are likely in any field of research in its early development, leading to differing research agendas (Andreasen, 1994) and, perhaps, slowing down incremental knowledge acquisition. In the case of the concept of 'ethnicity', unquestionably at the root of ethnic marketing, it is conceptualised relatively explicitly by Weber's (1961) as "a sense of common

descent extending beyond kinship, political solidarity vis-a-vis other groups, and common customs, language, religion, values, morality, and etiquette, (that) provides 'a set of sociocultural features that differentiate ethnic groups from one another' (Cohen, 1978:385), arguably reducing the scope of ethnic marketing and the contributions identified as ethnic marketing research. Yet, it is not uncommon for the concepts of ethnicity and ethnic identity to be afforded parity with a variety of consumer identities (e.g. tribes and communities) involving groups of people who are geographically dispersed but share some commonality, such as the Internet (Henderson et al., 2013) or sexual preference (Parks et al., 2004). While these groups of people are sometimes referred to as cultural groups, they are difficult to reconcile with ethnicity conceptualisations, hence to be qualified as ethnic. In the case of ethnic marketing, Pires and Stanton (2015) propose a rigorous definition expressed as "*the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for ethnic identified customers, clients, partners and communities, and for society at large*" (p. 9); however, a universally accepted definition is still lacking in the literature.

This discussion does not purport to merely advance critical perspectives on the way that ethnic marketing literature is commonly identified. Rather, the research purports to establish the dimensions of the literature that is currently identified under the ethnic marketing label. It is proposed that the analysis provides some support to perspectives regarding the undesirable conceptual latitude of 'ethnic marketing' recurrent in the literature, which is argued to inhibit the development of ethnic marketing research (Pires & Stanton, 2015).

2.2. Literature analysis and synthesis

The vast quantity of publications disseminated by numerous publishers with online presence and available through the internet poses an often challenging task when selecting the relevant

sources on a research subject (Fink, 2013). Online tools provide a means to narrow down the search, but the number of articles collected may still be very large, perhaps excessively so, making it difficult to ascertain the literature dimensions, and hard to conduct a proper literature analysis (Delen & Crossland, 2008). While manual coding is still a commonly adopted technique, the time frame involved increases linearly with the volume of literature under analysis (Sacha et al., 2016). Furthermore, the subjectivity inherent to human coding poses a difficult challenge in assuring the independence of the final result, usually requiring a panel of independent reviewers to address such issue (Ritchie et al., 2011). This justifies the usage of automated techniques for qualitative literature analysis of the selected literature such as automated textual analysis systems especially suited for large volumes of data. Classical content analysis, word count and automated text data mining are three of 17 methods of qualitative analysis identified by Onwuegbuzie et al. (2012).

Text mining allows the discovery and retrieval of high quality new semantic information by means of an automatic analysis of patterns and trends in the recurrence of text within the literature being analysed, while providing a deeper understanding of the contents than a simpler word count analysis (Gupta & Lehal, 2009). Drivers for conducting automated text mining include the volume of relevant literature, its availability in electronic format, and the availability and ease of use of text mining tools (Delen & Crossland, 2008).

The text mining approach for automated literature analysis uses an information systems application for parsing together a set of documents and for extracting its textual contents into an organized structure (the document term matrix) that encompasses two dimensions: the words (or terms, composed of n-words); and the documents. At the intersection of each pair of word and document (i.e., within each cell of the matrix) it is possible to retrieve the frequency of appearance of a word in the document (Delen & Crossland, 2008). Importantly, complementing text mining with a literature synthesis condenses the body of knowledge into

a manageable format that enables critical interpretation. This makes it more suitable to drive a critical discussion on the summarized literature, and to unveil insights on the subject under study (Schryen, 2015).

Several techniques have been developed in recent years for building knowledge models based on the topics extracted using text mining procedures. Two of the most common techniques are latent semantic analysis and topic modelling. The former is a natural language processing technique that analyses relationships between textual terms and documents by assuming that words that are close in meaning will occur in similar pieces of text (Dumais, 2004), while the latter uses the structure resulting from text mining as an input, with the relevant words and their frequency collected into an organized structure where the documents are distributed into topics (Blei, 2012). Both techniques produce topics which summarize the body of knowledge embedded in the documents, constituting a synthesis of the literature.

Applied to topic detection and spam filtering, Lee et al. (2010) performed a comparative analysis of four techniques for summarizing the knowledge hidden within textual sources, including two topic modelling techniques (the latent Dirichlet allocation and correlated topic modelling), the latent semantic analysis, and the probabilistic latent semantic analysis, which is a variant that uses a generative latent class model to perform a probabilistic mixture decomposition, instead of performing a singular value decomposition (Hofmann, 2001). The study found probabilistic latent semantic analysis to be the most accurate, followed closely by the latent Dirichlet allocation. Notably, the latent Dirichlet allocation can be best applied to documents dealing with multiple topics, a common characteristic of academic literature versing on a main subject, but also embedding other lines of research that a literature analysis needs to unfold for a deeper understanding of the themes involved within each publication. This justified adopting the latent Dirichlet allocation for the present study.

2.3. Ethnic marketing literature analysis and synthesis

There is evidence of a growing scholarly interest on "ethnic marketing", "multicultural marketing" and "cross-cultural marketing" as separate fields of research, with approximately 70 percent of all Google Scholar search results for each of the three fields occurring in the period from 2005 to 2015 (approximately 80 percent in Scopus). In the case of "ethnic marketing", querying Google Scholar with the specific two words side-by-side resulted in 1240 hits (on 9 March 2016, 859 of which in the period 2005-2015). On that date, the same query on Scopus resulted in 35 hits (29 in the period 2005 to 2015).

The relatively low number of hits emphasizes the contribution of the present research in identifying, analysing and synthesizing the dimensions of "ethnic marketing" research, a contribution also highlighted by the scarcity of literature offering reviews of the ethnic marketing field. Table 1 lists four articles offering analysis of ethnic marketing literature that could be missed in a superficial literature search. None of the articles refers to "ethnic marketing" in the title. Indeed, one of the articles refers to an assessment of cross-cultural marketing theory and research (Engelen & Brettel, 2011), illustrating the conceptual dissonance in the literature, identified earlier in this paper.

Table 1 – Analysis of ethnic marketing literature

Reference	Search	Method	Subject
Burton, 2000	Not specified	Critical review	Ethnicity, Identity and Marketing
Cui, 2001	Not specified	Critical review	Marketing to ethnic consumers
D'Rozario & Bryant, 2007	Not specified	Critical review	Differences within Immigrant / Ethnic Populations
Engelen & Brettel, 2011	1990-2008, 99 articles, 14 journals	Content analysis and classification	Cross-cultural marketing theory and research

Three of the articles in the table offer critical reviews. One involves a content analysis and classification for synthesizing a set of 99 articles appearing in 14 different journals. None of the articles used text mining as a technique for assessing the literature.

3. Approach and methods

3.1. Selecting articles

Most core academic literature is published in peer-reviewed journals, with the most relevant being indexed in both Web of Science (WoS) and Scopus (Moro et al., 2015). Since Scopus covers more titles, with a larger database (Abrizah et al., 2013), it was chosen to query for ethnic marketing articles in the current literature analysis. The purpose is to cover a substantial number of recently published literature on ethnic marketing, involving the gathering of articles published from 2005 to 2015, a period of 11 years. The result yielded a large sample of 469 articles, from 111 publishers and 310 different journals. Such range of publication titles is a reflection of the diversity associated with the thematic involving ethnic marketing issues and a variety of industries, from health (105 articles) to food and sports. Two extra filter selection procedures (filters A and B) were subsequently executed to improve the quality of the results and to reduce the size of the sample of selected articles.

Filter A was first applied, attending to publisher type and focusing on nine widely regarded publishers, shown in Table 2, which made up around 65% of the original Scopus query sample. Seven publishers are perceived as generic in terms of their coverage of the target sciences, as in the cases of Elsevier and Wiley. The two other publishers were selected for their specific focus on health and medicine (BMC - BioMed Central) and marketing (AMA - American Marketing Association). BMC was included given the large number of health

related manuscripts focusing on ethnic marketing issues. AMA is a well-established marketing association.

Filter B consisted of a final manual inspection, seeking to establish the true relevance of the sample units within the ethnic marketing domain. This was achieved by examining the titles and abstracts of the 306 papers that resisted to the filter A scrutiny. For example, Richardson (2015) was removed for lack of relevance. Published by Wiley, the focus of the paper is on post-genomic research, hence does not relate directly with marketing and ethnicities, even though it mentions both terms (e.g., in the Abstract: *"…work on racial and ethnic variation in brain… the possibility of marketing prenatal tests…"*).

Application of filter B resulted in 67 articles being discarded, with the remaining 239 articles being retained and subject to the text mining analysis approach. This set of 239 articles is briefly characterised in tables 2, 3 and 4.

Table 2 – List of Selected Publishers

Publisher	Number of articles		
	Initial	**Removed**	**Retained**
Taylor and Francis	82	9	73
Elsevier	74	18	56
Emerald	47	8	39
Wiley	31	16	15
SAGE	30	2	28
Springer	21	6	15
BioMed Central	13	8	5
American Marketing Association	6	-	6
M. E. Sharpe	2	-	2
Total	*306*	*67*	*239*

Table 2 shows the number of articles removed and retained for each publisher. Somewhat surprisingly, eight articles published by BioMed Central, were eliminated from the original dataset, perhaps due to its focus on medicine. Shifting attention to the journals where the retained articles are published, Table 3 shows a heterogeneous distribution, with only seven journals carrying four or more relevant publications.

Table 4 shows the number of articles in the sample distributed by year of publication. It is apparent that the yearly number of publications more than doubled in 2007, remaining mostly steady for the rest of the period, with the exception of a peak of 33 publications in 2009.

Table 3 – Journals contributing the most articles

Journal	Publisher	Number of articles
Journal of Consumer Marketing	Emerald	8
Journal of Business Research	Elsevier	6
Journal of Travel & Tourism Marketing	Taylor and Francis	5
Journal of Services Marketing	Emerald	5
Preventive Medicine	Elsevier	5
American Journal of Preventive Medicine	Elsevier	5
Journal International Consumer Marketing	Taylor and Francis	4

Table 4 – Distribution of articles from 2005 to 2015

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
No.	15	10	21	22	33	22	19	26	26	23	22	**239**

With 33 articles, 2009 is a clear outlier and was examined separately. It was found that four articles relate to sports involving ethnicity issues. But the majority of the difference (11 more

articles in 2009 relative to 2008) refers to 15 articles dealing with health marketing issues, (only three in 2008). Table 5 lists the subjects addressed by these 15 articles.

Table 5 – Health marketing subjects for the articles published in 2009

Health subject addressed	Number of articles
Alcohol	1
Food	3
Obesity	4
Smoking	3
Stress	1
Supplements & medications	3
Total	**15**

Obesity accounts for four articles, while food, smoking and supplements & medications account for three each, with all of them having a relation to health marketing issues, and mentioning both "ethnic" and "marketing" on the indexed sections. One example of such articles is the study of a social marketing campaign to promote folic acid use among Hispanic women, by Quinn et al. (2009), hence classifiable on the supplements & medications subject.

3.2. Text mining

Text mining may be used for a blind search of every word in specific sections of articles in the hope of revealing insightful trends (Delen & Crossland, 2008). However, the result typically demands a dimensionality reduction that produces a smaller set of documents for consideration, capable to provide a manageable structure that contains all findings (Fortuna et al., 2005). This reduction might be achieved by building a dictionary of relevant terms on the specific subject of interest, using it to guide the text mining procedure through the articles' contents (Moro et al., 2015), effectively filtering and purifying the set of documents being

considered. This is the procedure adopted in the present analysis, although using two dictionaries.

With the purpose of building the final set of ethnic marketing related documents to be analysed and synthesised in the present research, two distinct but related dictionaries were devised for scrutinising the contents of the selected articles:

- A "**sufficient dictionary**", containing the terms whose occurrence in a document makes it sufficient for the document to be identified under the label "ethnic marketing"; and,

- A "**conditional dictionary**", containing a larger number of terms (such as minority, cultural group, community) that, suggesting a possible relationship to the topic of interest, may or may not be actually related to ethnic marketing research; the decision about this relationship is conditional to a clear identification of the document with keywords from the "sufficient dictionary".

The "sufficient dictionary" was developed based on the identification of the most meaningful words in the following definition of "ethnic marketing" formulated by Pires and Stanton (2015:6; 55-56):

*"... the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for **ethnic identified** customers, clients, partners and communities, and for society at large (AMA, 2007). To be justified when focused on **ethnic minority** consumers, these consumers need to constitute an effective and viable market, offering an opportunity that requires a market definition grounded on **ethnicity** as the basis for the market segmentation decision."*

This definition encompasses a restricted set of keywords, including those underlined in the above quotation and consistent with the derivation of "ethnocentrism", defined by *The Free Dictionary* as an "overriding concern with ethnicity" (http://www.thefreedictionary.com/ethnocentrism).

The "conditional dictionary" includes several terms that recur in the marketing literature but that can also occur in contexts unrelated to ethnic marketing. These terms by themselves are not effective indicators of an ethnic marketing related document. The terms in the dictionary used in the present study were endorsed by a panel of experts on ethnic marketing literature, assessed by their recurrence in documents related to ethnic marketing and/or ethnicities as categorised by Huntington (1997). The list of terms in the conditional dictionary was then analysed to determine whether they occurred in the title or in keywords identified within the documents. The aim was to verify that the terms identified by the panel could be expected to appear in at least one of the two most relevant sections in each article, for indexation purposes.

Both dictionaries were subsequently merged together to assure an indiscriminate modelling procedure (Moro et al., 2015). The result is displayed in **Error! Reference source not found.**6. It should be noted that only the reduced terms are shown for simplification purposes, although stemming was applied to account for word variation, including distinct words with similar meaning, such as the ethnicities, as defined by Huntington (1997). Also, it should be stated that ethnolinguistic was included in the conditional dictionary since it is defined as "the study of how language relates to culture and ethnicity" (Underhill, 2012, p. i), thus the relationship to culture deviates from the core subject related to ethnic studies, even though the prefix of the word may suggest otherwise. On the opposite, "ethnocentrism" is a "psychosocial construct with relevance to individual-level personality systems as well as to the more general cultural and social analytic frameworks" (Shimp & Sharma, 1987).

Thus, it is at the core of ethnic marketing studies, justifying its inclusion within the sufficient dictionary.

Table 6 – Ethnic marketing dictionary.

Sufficient dictionary

ethnic	ethnicity	ethnocentrism

Conditional dictionary

acculturation	human resources	religion	African
communication	identity	sensitivity	Buddhist
contradictory effect	intensity	service-dominant	Hindu
conventional method	issues	skepticism	Islamic
decision process	market segmentation	social responsibility	Japanese
demography	marketing theory	strategy	Latin American
discrimination	minorities	substantiality	Orthodox
disparities	opportunities	suppliers	Sinic
diversity	participation	three-dimensional construct	Western
drivers and future	politics	tourism	
ethnolinguistic	prognosis	transnationality	
globalization	relational perspective		

 * All terms are in lower case and separated by commas.

The result of the text mining procedure on a document is a term matrix, which is a data structure with two dimensions, the relevant terms, and the documents where each relevant term occurs, with each cell of the matrix containing the frequency of occurrence (Lui et al., 2007). Such structure, exemplified on Table 77, can then be used as an input to data mining algorithms for extracting useful knowledge that synthesizes the collected documents. Merging all terms in a single set assures that all have the same value as an input for the topic modelling algorithm, whether a term is sufficient or conditional for the study being addressed. Thus, the algorithm will be agnostic to this division. It will only be in the moment

of the interpretation of the topics discovered through modelling that the results will be able to shed some light on how the literature is being classified under the label "ethnic marketing".

Table 7 – Example of a document term matrix

Article reference	ethnic	Latin American	minorities
Baumann and Setogawa (2015)	190	0	13
Schuft and Massiera (2012)	42	0	0
Yancey et al. (2009)	21	47	3

For example, Moro et al. (2015) merged two distinct dictionaries, one for Business Intelligence terms and one for banking terms, prior to proceeding to identify the three most relevant terms per topic, indistinctly from the two source dictionaries. As a result, four from the nineteen topics modelled "*were best identified by three terms, all related to Business Intelligence*" (p. 1321), leading to the conclusion that a significant collection of articles used banking problems as benchmark for evaluating cutting edge Business Intelligence techniques, not clearly evaluating in-depth benefits to banking. On the same line of thought, by evaluating the terms that best characterize each topic on ethnic marketing, it is expected that some topics can be identified as clearly related to ethnic marketing, while other emerging topics are representative of the misclassification of the literature on ethnic marketing described in the earlier sections.

3.3. Approach for literature analysis and synthesis

Figure 1 provides a visual scheme of the proposed approach. Relative to the methodology prescribed by Moro et al. (2015), two major improvements are introduced:

1. The definition of the dictionary adopts a 3-step generalizable procedure (described in Section 3.2 for the ethnic marketing context), based on defining:

    i. A sufficient dictionary that contains the terms that unequivocally identify the relevancy of a document to the theme being studied. This dictionary is based on a standard theme definition, and involves identification of the main keywords;

    ii. A conditional dictionary, encompassing the terms that usually appear in documents related to the theme being examined but that, by themselves, are not sufficient to assure a matching of the document to the theme. This dictionary is based on a mixed approach constituted by an initial expert panel assessment and literature definitions. Then, the dictionary is validated through comparison with indexed sections of the documents, assuring term congruency;

    iii. The final step consists of the merging of the two dictionaries, which results in a unique dictionary to be used for feeding the text mining and topic modelling procedures.

2. The adoption of a visually appealing scheme based on a thematic concept map that identifies the topics that match the main research theme (an indication of relevancy), and the topics that do not, hence providing a visual interpretation of mismatches in the literature. A word cloud is also defined per topic using the set of documents that are deemed relevant to the topic, to allow an interpretation of the most frequent terms and help supporting the interpretation from the characterization of each topic (Heimerl et al., 2014). This is also a differentiating feature from Moro et al. (2015), as it is focused on producing a word cloud per topic, while the cited paper provided a word cloud including every document.

The proposed approach comprises three distinct sections: a literature collection - to retrieve the relevant documents on the subject; a literature analysis - to gather all the relevant information from the documents; and a literature synthesis - to summarize the results in comprehensible topics. Each of these is discussed below.
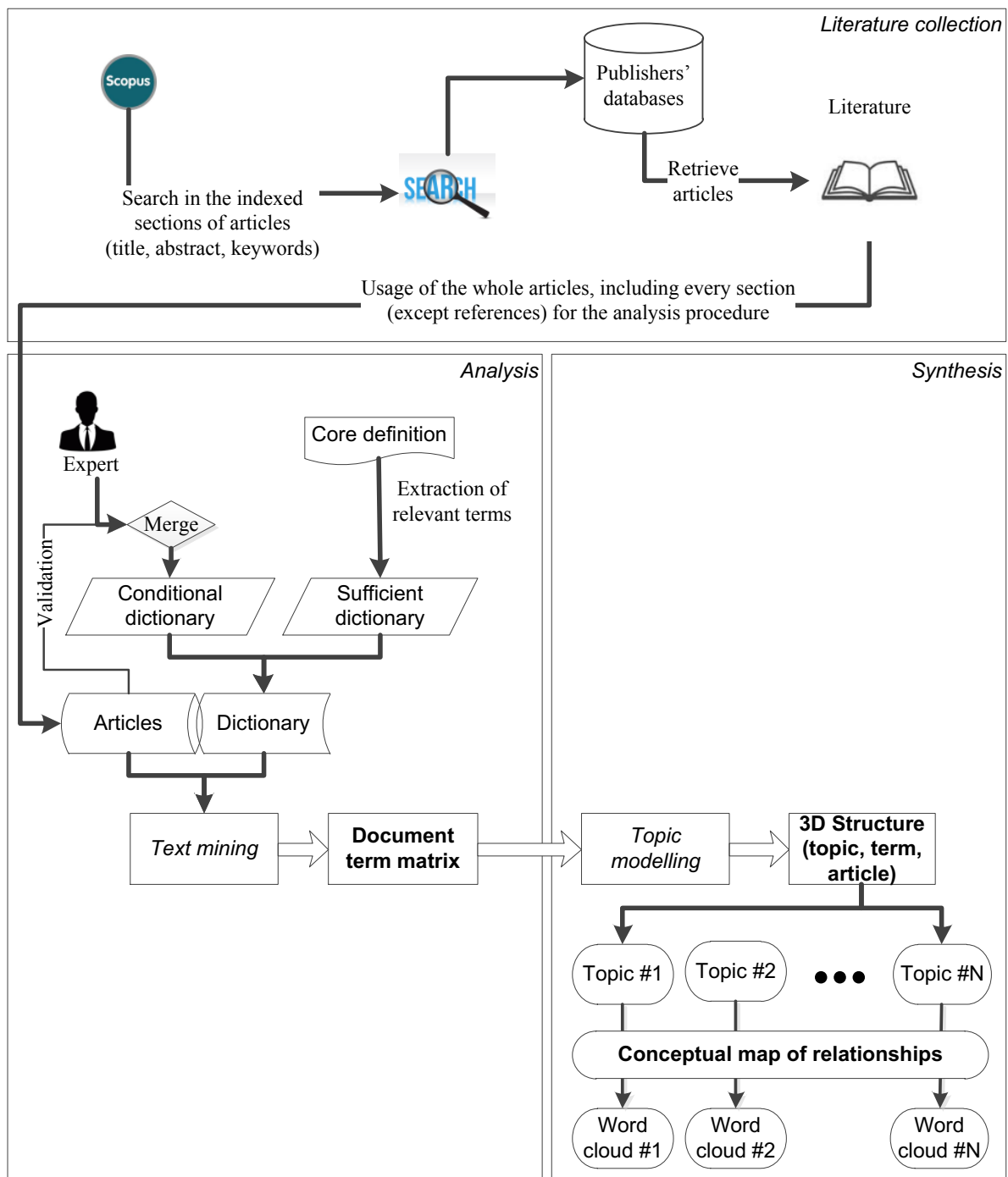

Figure 1 – Proposed approach.

The literature collection uses Scopus, an academic search engine that indexes each article's title, abstract and keywords. The search is focused on selecting only the literature that mentions the query words on the most relevant parts that summarize each article. Then, the selected articles are retrieved from each publisher, for building the corpus. Finally, every article's entire textual content is used for feeding the literature, excluding the references, which may contain words from the dictionaries in the titles of cited articles, possibly misleading the proposed procedure. As argued by Moro et al. (2015), *"the full-text analysis allows a better evaluation of term frequencies, since a term expressed numerous times through an article is probably more relevant than another that is only mentioned in the abstract"* (p. 1318).

The literature analysis adopts a text mining approach to collect relevant terms from each document and to organize the resulting information in a term matrix structure. Applied to the ethnic marketing context in this paper, this matrix is a large bi-dimensional structure with 239 articles and 47 terms, corresponding to 112,339 cells that reflect each relevant term's frequency in each document. While Moro et al. (2015) analysed the entire lexicon gathered through text mining, we proceed here with a synthesis procedure using topic modelling. The rationale behind this change is that a glimpse on the most frequent terms for all the collected literature contributes little to answer the research question of whether the papers collected through a query on the sufficient terms are classifiable as belonging to the ethnic marketing discipline.

The advantage of using specific domain dictionaries is highlighted by Stanfill et al. (2010) as applied to critical clinical systems: context knowledge is a key asset in assuring the proper behaviour of such systems and must be included to improve systems' accuracy. Following this line of thought, using dictionaries helps to cope with the challenge of finding the right patterns of information, as stated by Moro et al. (2015). However, contextual knowledge

often comes at the cost of subjectivity. Thus, in order to reduce the impact subjectivity, using a standard definition of the theme becomes an important improvement when compared to the previous study by Moro et al. (2015), which defined the dictionary based on a panel of experts' opinions.

The literature synthesis is an attempt to summarize the results from the large term matrix in a few topics that condense the ethnic marketing literature and from which trends may be inferred, helping to identify research gaps for the ethnic marketing discipline. A few known algorithms may be used for this topic discovery, the latent Dirichlet allocation brings the most popular topic modelling technique, commonly used for a wide variety of problems (Blei, 2012), including literature analysis (Moro et al., 2015). This is also the algorithm adopted for the procedures undertaken in this research.

The topic modelling procedure takes as input the document term matrix that is subsequently subject to an algorithm designed to find correlations between terms and documents. This results in a tri-dimensional structure characterized by relationships between each of the topics, documents and terms. Thus, for any given document, it is possible to determine how it relates to any topic or to any term, characterized by a beta (β) distribution (with a β value closer to zero representing a stronger relationship).

The output from topic modelling can be presented in a visually appealing picture based on a conceptual map (Maher & Fisher, 2012). The map shows the topics ordered by the most to the least related to the main theme (ethnic marketing, as characterized through the "sufficient dictionary"), displaying also the conditional subjects related to ethnic marketing. For a deeper analysis of each topic, word clouds are drawn from the articles that are most related to each topic, allowing the provision of a visual interpretation of the frequency of all terms, emphasizing the most frequent terms through a larger font size (McNaught & Lam, 2010).

The number of topics is an exogenous parameter that must be provided to the modelling algorithm. Therefore, the number was first set to twenty, and then, following Moro et al. (2015), it was reduced through a series of consecutive experiments until a reasonable level of aggregation of the documents in topics was achieved. Eight topics were chosen to summarize all the literature considered.

The experiments were conducted using the R statistical tool, specifically both the "tm" and "topicmodel" packages. The latter package can be directly fed with the document term matrix produced by the former to discover topics through the latent Dirichlet allocation algorithm adopted for the experiments. In order to apply the dictionaries to the corpus of documents, specific code in R was developed, mainly to read the dictionaries and to build a lookup table for direct use during the analysis stage.

4. Experiments and discussion of the results

The experiments conducted by applying the methodology illustrated in Figure 1 seek to summarize the literature collected on ethnic marketing, addressing the research question about the dimensions that characterize ethnic marketing in the literature published in the 2005-2015 period. Pictorial representation of the findings allows each topic's proximity to the sufficient terms that characterize ethnic marketing to be visually interpreted. By also displaying the terms from the conditional dictionary, it is possible to ascertain the terms that are not directly relevant to ethnic marketing, indicating that the "ethnic marketing" label may be inappropriately applied to some of the examined literature.

Figure 2 displays the eight topics defined by the latent Dirichlet allocation algorithm in vertical columns separated by dots. The vertical axis defines the β distribution value with a closer value to zero representing a stronger association; the rounded squares with black background are positioned according to the β distribution value of the proximity to the

sufficient dictionary. Hence, the topics are numbered from the strongest to the weakest association with ethnic marketing. Each rounded square shows the number of documents with strongest relevancy (the best match) to the topic. Each topic also shows the terms belonging to the conditional dictionary ($\beta < 2.5$), positioned in respect to the corresponding β value. This allows evaluation of the remaining dimensions.
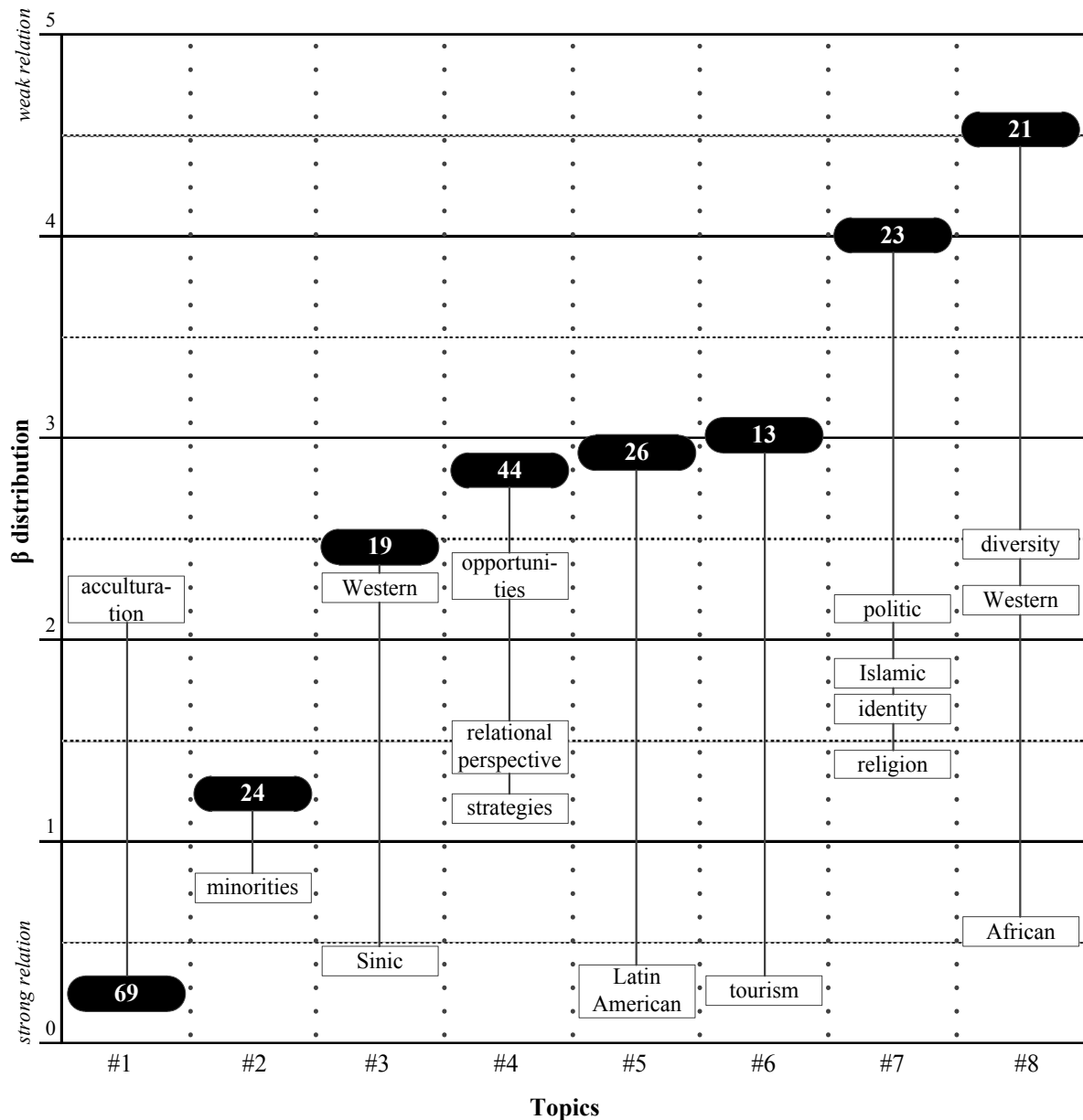

Figure 2 – Characterization of the topics discovered.

The first observation (topic #1) from the characterisation of topics in Figure 2 is the only one topic highly associated with terms in the sufficient dictionary for ethnic marketing. While this is by far the largest topic with 69 related documents, corresponding to approximately 29% of the 239 articles collected (the second largest has 44 articles and the third only 26), its uniqueness and relatively low frequency suggests that much self-identified ethnic marketing literature does not fit the definition forwarded in section 2.1.

Alternatively, the characterisation may be interpreted to indicate the lack of a clear, widely accepted definition of ethnic marketing in the literature. The association of Topic #1 to "acculturation" is not strong ($\beta > 2$). In fact, the word cloud displayed in Figure 3 for this topic reveals that the term "acculturation" does not distinguish much more than other conditional terms such as 'identity" or "relational perspective", indicating that topic #1 encompasses a subset of the literature directly related to ethnic marketing. The conditional subjects that stand out the most besides acculturation include identity, relational perspectives and minority, with ethnicities Latin American, Sinic, African and Western also deserving a significant share of attention under the ethnic marketing umbrella. The Western ethnicity emerges from two topics, both #3 and #8, which may be caused by the historical influence the Western world had in the remaining societies, i.e., Sinic, and African countries. Additionally, it may also be due to research historically focused on Western countries, or on comparisons to Western countries (Australia, the US, the UK, and Canada) (Huntington, 1997).

All these themes have been considered in the specialist literature as drivers or influencers of ethnic marketing (Pires & Stanton, 2005). Such result shows that scholarly devotion to these themes has not changed much since, suggesting the possible need to envision the incorporation of influencing factors in a world where acknowledging the difference may stand out, leveraging businesses worldwide.
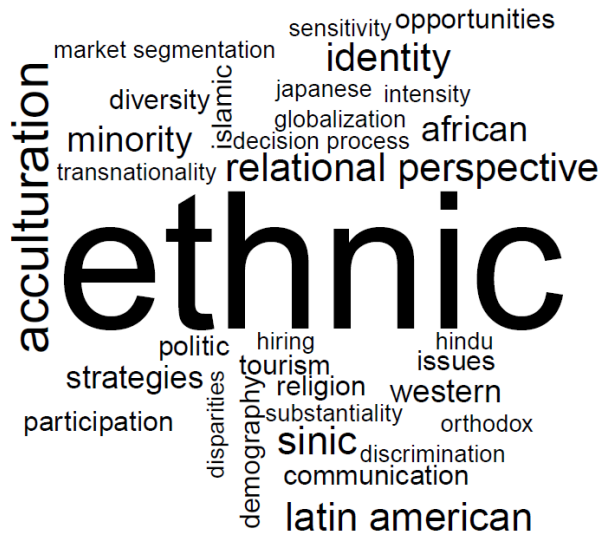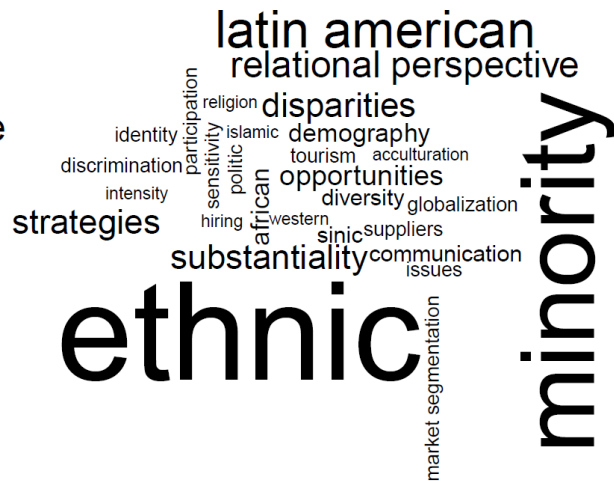
Figure 3 – Word cloud for topic #1.　　**Figure 4** – Word cloud for topic #2.

Topic #2 with 24 documents reveals a tight association between ethnic marketing and the term 'minorities' (with both terms separated by around 0.5 of difference in the β distribution). This is also indicated in the word cloud representation in Figure 4, which shows the sufficient term "ethnic" slightly larger than the term 'minority'. It should be noted that the word cloud is a direct visualization of the frequency of terms occurring in the 24 articles that are most related to this topic, while some of these articles may also have relationships with other topics. However, the topic is more strongly associated with "minorities" than with ethnic marketing. Such finding is reflected in the articles directly related with this topic, such as Grier and Davis (2013), which focuses on understanding the relationship between retail marketing and weight among adolescents across urban ethnic minorities. A prescription that arises from this topic is the need to clarify within each study on minorities if the focus is on ethnic minority groups (e.g., Pires & Stanton, 2002) or on other types of minorities related to a vast array of themes such as gender, wealth, health or sexual orientation. Such differentiation could lead future ethnic marketing research to more focused clusters of distinct subjects.
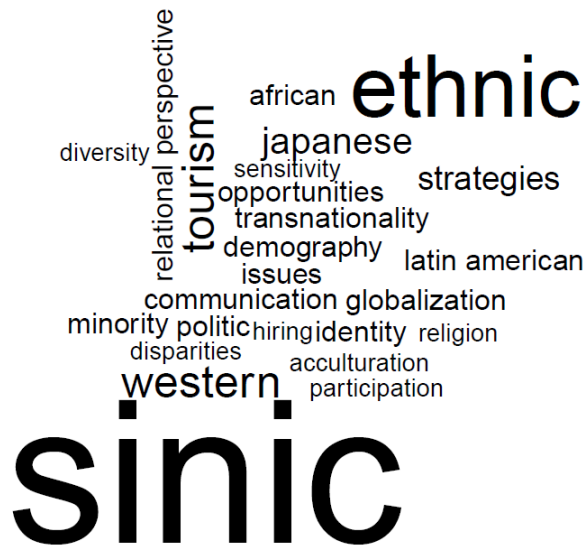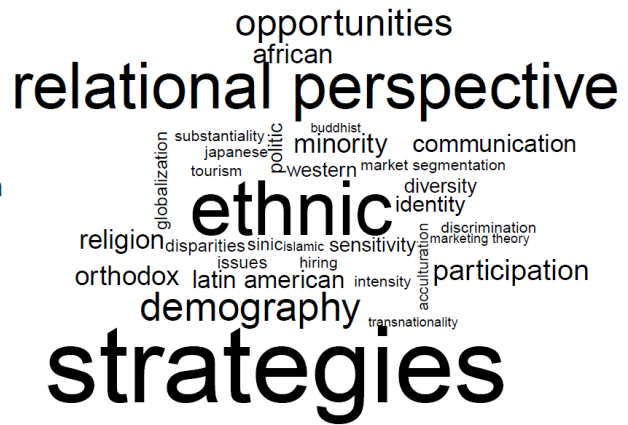
Figure 5 – Word cloud for topic #3.

Figure 6 – Word cloud for topic #4.

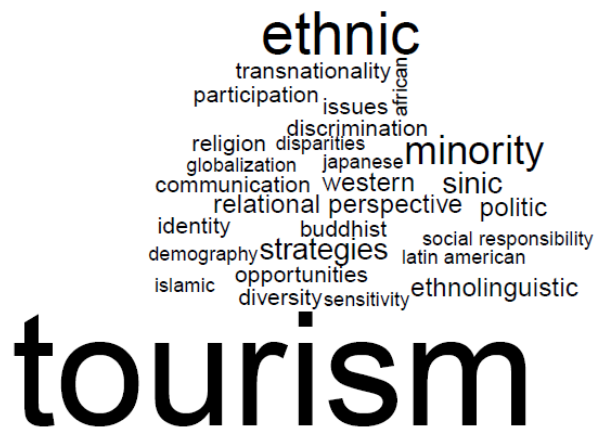**Figure 7** – Word cloud for topic #5.

Figure 8 – Word cloud for topic #6.

Topics #3 to #6 are all lightly associated to ethnic marketing, having much stronger associations with other conditional terms. Topics #3 and #5 show strong associations with two ethnicity terms, respectively "Sinic" and "Latin American" (word clouds in Figure 5 and Figure 7). Topic #4 is the second largest with 44 documents, revealing its cross-subject and

multi-ethnicity nature, as shown in Figure 6, with an emphasis on terms such as "strategies" and "relational perspective", and with less relevance for "opportunities" and "demography". Topic #5 shows a strong association with "tourism" (Figure 8). Tourism marketing is an active term in many domains, including ethnic and minority issues, among others (e.g., Min & Lee, 2014). While the four topics include articles that briefly mention ethnic marketing related issues, most of the literature encompassed orbits around other main themes.

Given its characteristics, ethnicity may be regarded as a different tree of research, with each ethnicity being a major branch. Therefore, specific studies accounting for such characteristics are in demand, whether in the domain of marketing, or human resources management, tourism and hospitality, among others. Specific studies grasping onto the conspicuous characteristics of ethnic marketing may help to shed some light on marketing related issues targeting one specific ethnicity, contributing to ethnic marketing knowledge, while other studies may be seen as more related to ethnicity management issues. Examples of the latter case include the study by Stowell et al. (2012) versing on Latin American crime related to Latin ethnic groups, from topic #5, and the study by Sim Ong et al. (2008), aiming at identifying consumption patterns in Malaysia, including specific ethnic groups, from topic #3.
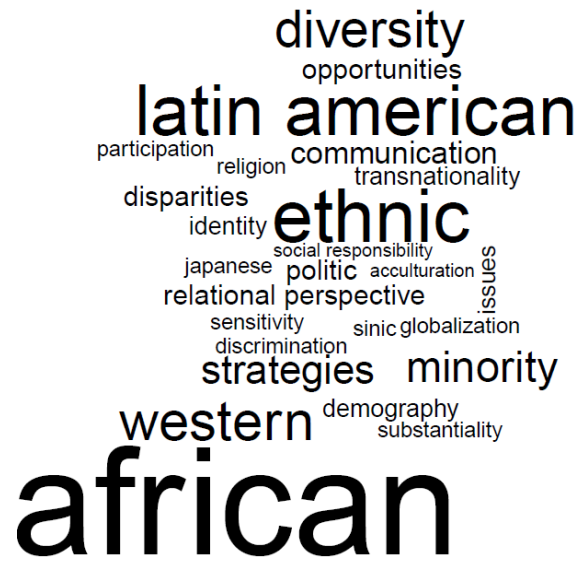
Figure 9 – Word cloud for topic #7.   **Figure 10** – Word cloud for topic #8.

Finally, topics #7 and #8 have weak associations with ethnic marketing and include such distinct terms as "Islamic", "Western" and "African" ethnicities (Figure 9 and Figure 10). Together, these two topics encompass 44 highly associated documents, providing clear support for the argument in Section 2.1 that there is an urgent need to purify the ethnic marketing literature by clarifying its dimensions. These two topics can hardly be seen as ethnic marketing studies, being more related to the context associated with specific ethnicities ("Islamic" and "African"), along with the already mentioned topics #3 and #5.

5. Conclusion

5.1. Contributions to ethnic marketing research

Cited earlier in this paper, Pires and Stanton (2015) definition of ethnic marketing applies the AMA endorsed definition of marketing, available at https://www.ama.org/AboutAMA/Pages/Definition-of-Marketing.aspx to environments exclusively defined by ethnicity. Rather than subjecting that definition to unjustified scrutiny,

since it remains undisputed, this paper contributes to ethnic marketing research by unfolding a high level of conceptual heterogeneity apparent in the intrinsic nature of articles published under the ethnic marketing umbrella, yet clearly incongruent with that definition. It is this incongruence that may inhibit a more grounded development of specialised research focusing on improving marketing actions specifically designed for targeting ethnic groups. Hence, the suggestion is that, instead of prescribing a new conceptualisation and definition of ethnic marketing, this paper calls for research that carefully and explicitly considers the underlining view of ethnic marketing as involving ethnicity based segmentation vis. a vis. other possible interpretations given to the "ethnic marketing" expression, such as "multicultural marketing" or "cross-cultural marketing".

The unprecedented literature analysis conducted in this paper reveals the need for a reengineering of "ethnic marketing" that removes any ambiguity in what is at the core of such specific branch of marketing and what are the satellite research themes that orbit on this sphere. Furthermore, defining the borders of such sphere of knowledge may shed some light on the valuable work of numerous researchers worldwide, who investigate seemingly related subjects but lying outside of this sphere, such as marketing to minority groups defined by religion or lifestyle. The biasing of future research based on ethnicity is expected to lead to a more fruitful indexation of better defined titles, abstracts and keywords of forthcoming articles, as well as providing a solid ground for incremental knowledge on ethnic marketing.

5.2. Contributions to literature analysis in the age of big data

The literature analysis method espoused in this paper is not meant to substitute for critical analyses of the literature. In its present state, the method neither allows for different arguments on the same topic to be discerned, nor does it support qualitative assessments of the literature. Notwithstanding, the method offers extensive potential for delimitating the

literature, identifying main themes and key words, as well as potential multidisciplinary overlaps and rigidities in research, hence facilitating potential remedial actions.

An important aspect to emphasise is the efficiency of the method comparatively with the performance of an in-depth critical analysis, especially in the present age of big data, with numerous sources of online information ready to be analysed. Therefore, the procedure devised can precede a critical analysis by pruning less relevant papers.

This paper sought to develop a holistic method to increase both the reliability of the dictionary and its applicability to complex multidisciplinary research domains, as demonstrated in its application to ethnic marketing. The definition of sufficient and conditional dictionaries is a new contribution that allows a more multi-faceted analysis than it was possible to date, allowing a better understanding of the trends, and eventual problems, in the relevant literature. This is extensive to longitudinal analysis, depending on whether the focus is on a given period of time (2005 to 2015 in the present case), or whether the analysis distributes the literature over different time intervals.

This paper also proposed that the sufficient dictionary can and should optimally be defined based on a well-established, universal definition of the subject of interest, eliminating the subjectivity associated with user-defined dictionaries. This same issue was addressed for the conditional dictionary, with identification of terms by an expert panel, then cross-checked against each of the articles' title and keywords.

A further contribution of this paper lies in the reporting method that is applied, which involves the presentation of the topics in a novel visually appealing picture (topic map) that drastically increases readability and interpretability when compared to the complex tables use in the extant literature. It should be recalled that the topic modelling structure is a complex

tri-dimensional matrix characterized not only by the topics, but also by the terms and the documents included in the corpus analysed.

The improvements presented on the analysis method can be viewed as a stepping stone toward scalability. The adoption of more precise procedures for defining the dictionaries without relying solely on expert human knowledge facilitates the analysis of a larger body of documents, whereas previous method required a manual reading of the indexing portions of the articles, posing a time consumption task as the corpus grows larger. At the same time, by using a simpler visual map for showing the topics discovered, reporting eases the burden of analysing a large number of topics that may arise with a larger corpus. These are important contributions, especially contextualising this method in the age of big data, with vast quantities of scientific sources and papers to be searched for a particular research subject.

5.3. Limitations and future research

While the method forwarded in this paper is developing and has been applied in the recent literature for topic modelling in different contexts, this paper strengthens one of the major methodological limitations, relative to the selection of the literature to be analysed before subsequent synthesis and topic modelling can proceed. This stepped sequence reveals the dependency on the quality of the literature filtering tool, the dictionary. But improving the dictionaries still depends on the quality of the terms identified by the expert panel. While this is acknowledged as a major area for further research, it does not appear to compromise the suitability of the methodology for the purpose that is being devised. Notwithstanding, the method calls for continued refinement of dictionaries, which need to be subject to more testing.

While recognising the relevance of expert domain knowledge in extracting the appropriate information from the text mining procedures, this is a process that is still amenable to bias,

and problematic when using big data. One approach for addressing such limitation would be to devise a clear process for the use of expert panels to create the dictionaries. Alternatively, an automated text mining approach for building up the conditional dictionary could be adopted, based only on the indexed sections of articles (testing if including the abstract would generate a too large dictionary and excluding abstracts as needed). The advantage of this latter approach relates to scalability, an important issue to consider when textual sources become big data.

References

Abrizah, A., Zainab, A., Kiran, K., & Raj, R. (2013). LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus. *Scientometrics*, 94(2), 721–40.

ABS (Australian Bureau of Statistics). 1249.0 - Australian Standard Classification of Cultural and Ethnic Groups (ASCCEG). (2011) [accessed 5.2.2016 at http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1249.0main+features22011].

AMA, Definition of 'Marketing", Approved October 2007. (2007). [accessed 11.3.2016 at http://www.marketingpower.com/aboutama/pages/definitionofmarketing.aspx.]

Andreasen, A. (1994). Social marketing. Its definition and domain. Journal of Public Policy & Marketing, 13(11), 108-14.

Armstrong, M. (2013). Minorities – the New Majority, Armstrong Economics, June [accessed 4.2.2016 at http://armstrongeconomics.com/2013/06/14/minorities-the-new-majority/]

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing*, 67(4), 103–17.

Baumann, C., & Setogawa, S. (2015). Asian ethnicity in the West: Preference for Chinese, Indian and Korean service staff. *Asian Ethnicity*, 16(3), 380–98.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.

Burton, D. (2000). Ethnicity, identity and marketing: a critical review. *Journal of Marketing Management*, 16(8), 853–77.

Cleveland, M., Laroche, M., & Papadopoulos, N. (2015). You are what you speak? Globalization, multilingualism, consumer dispositions and consumption. *Journal of Business Research*, 68(3), 542–52.

Cohen, R. (1978). Ethnicity: Problem and focus in anthropology. *Annual Review of Anthropology*, *7*(1), 379-403.

Cui, G. (2001). Marketing to ethnic minority consumers: A historical journey (1932-1997). *Journal of Macromarketing*, 21(1), 23–31.

Cui, G. (1998). Ethical Issues in Ethnic Segmentation and Target Marketing. In Proceedings, Chebat, Jean-Charles and Oumlil, A. Ben (eds.), 1998 Multicultural Marketing Conference (AMS, Montreal), pp. 87–91.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3), 1707–20.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1), 188-230.

D'Rozario, D., & Bryant, F. K. (2007). Re-conceptualization of literature in marketing on the measurement of differences within immigrant/ethnic populations. *Journal of Immigrant & Refugee Studies*, 5(4), 75–102.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, 14(4), 532–50.

Engelen, A., & Brettel, M. (2011). Assessing cross-cultural marketing theory and research. *Journal of Business Research*, 64(5), 516–23.

Fink, A. (2013). *Conducting research literature reviews: From the Internet to paper,* Sage Publications.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2005). Visualization of text document corpus. *Informatica* 29(4), 497–502.

Grier, S., & Davis, B. (2013). Are all proximity effects created equal? Fast food near schools and body weight among diverse adolescents. *Journal of Public Policy & Marketing*, 32(1), 116–28.

Guerreiro, J., Rita, P., & Trigueiros, D. (2016). A Text Mining-Based Review of Cause-Related Marketing Literature. *Journal of Business Ethics*, published online *25 March,* DOI 10.1007/s10551-015-2622-4

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60–76.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In System Sciences (HICSS), 2014 47th Hawaii International Conference on (pp. 1833–42). IEEE.

Henderson, G. R., Guzmán, F., Huff, L., & Motley, C. M. (2013). The Ian's pizza tribe: Reconceptualizing cross-cultural research in the digital age. *Journal of Business Research*, 66(3), 283–7.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1–2), 177-96.

Huntington, S. P. (1997). *The clash of civilizations and the remaking of world order*. Penguin Books India.

Jamal, A., Peñaloza, L., & Laroche, M. (Eds.). (2015). *The Routledge companion to ethnic marketing*. Routledge.

Lee, S., Baker, J., Song, J., & Wetherbe, J. C. (2010). An empirical comparison of four text mining methods. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on (pp. 1–10). IEEE.

Lui, A. K. F., Li, S. C., & Choy, S. O. (2007). An evaluation of automatic text categorization in online discussion analysis. In Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on (pp. 205–9). IEEE.

Maher, M. L., & Fisher, D. H. (2012). The role of AI in wisdom of the crowds for the social knowledge on sustainability. In Proceedings, AAAI Spring Symposium on Wisdom of the Crowd (pp. 44–9).

McNaught, C., & Lam, P. (2010). Using Wordle as a supplementary research tool. *The Qualitative Report*, 15(3), 630–43.

Min, K. H., & Lee, T. J. (2014). Customer satisfaction with Korean restaurants in Australia and their role as ambassadors for tourism marketing. *Journal of Travel & Tourism Marketing*, 31(4), 493–506.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, 42(3), 1314–24.

Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through datamining: The case of Las Vegas Strip. *Tourism Management Perspectives*, 23, 41-52.

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 69(9), 3341–51.

Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. (2012). Qualitative analysis techniques for the review of the literature. *The Qualitative Report*, 17(28), 1–28.

Parks, C. A., Hughes, T. L., & Matthews, A. K. (2004). Race/ethnicity and sexual orientation: Intersecting identities. *Cultural Diversity and Ethnic Minority Psychology*, 10(3), 241–54.

Pires, G. D., & Stanton, J. (2015). *Ethnic Marketing – Culturally Sensitive Theory and Practice*, London, Routledge.

Pires, G. D., & Stanton, J. (2002). Ethnic marketing ethics. *Journal of Business Ethics*, 36(1–2), 111–8.

Pires, G. D., & Stanton, J. (2005). *Ethnic marketing: Accepting the challenge of cultural diversity*. Cengage Learning EMEA.

Quinn, G. P., Thomas, K. B., Hauser, K., Rodríguez, N. Y., & Rodriguez-Snapp, N. (2009). Evaluation of educational materials from a social marketing campaign to promote folic acid use among Hispanic women: insight from Cuban and Puerto Rican ethnic subgroups. *Journal of Immigrant and Minority Health*, 11(5), 406–14.

Richardson, S. S. (2015). The Trustworthiness Deficit in Postgenomic Research on Human Intelligence. *Hastings Center Report*, 45(S1), S15–S20.

Ritchie, J. B., Tung, V. W., & Ritchie, R. J. (2011). Tourism experience management research: Emergence, evolution and future directions. *International Journal of Contemporary Hospitality Management*, 23, 419–38.

Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., ... & Keim, D. A. (2016). Visual interaction with dimensionality reduction: a structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics, in press*.

Schryen, G. (2015). Writing qualitative IS literature reviews–Guidelines for synthesis, interpretation and guidance of research. *Communications of the AIS*, 37(Art 12), 286–325.

Schuft, L., & Massiera, B. (2012). Marketing a nation by the performance of gendered, exotic bodies in sport and beauty contests: the case of Tahiti. *Sport in Society*, 15(1), 103–16.

Shimp, T. A., & Sharma, S. (1987). Consumer ethnocentrism: construction and validation of the CETSCALE. Journal of Marketing Research, 280-289.

Sim Ong, F., Kitchen, J. P., & Jama, A. T. (2008). Consumption patterns and silver marketing: an analysis of older consumers in Malaysia. *Marketing Intelligence & Planning*, 26(7), 682-98.

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6), 646–51.

Stowell, J. I., Martinez Jr, R., & Cancino, J. M. (2012). Latino crime and Latinos in the criminal justice system: Trends, policy implications, and future research initiatives. *Race and social problems*, 4(1), 31-40.

Underhill, J. W. (2012). Ethnolinguistics and cultural concepts: truth, love, hate and war. Cambridge University Press.

Upreti, B., Asatiani, A., & Malo, P. (2016). To Reach the Clouds: Application of Topic Models to the Meta-review on Cloud Computing Literature. In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS). Institute of Electrical and Electronics Engineers (pp. 10–21), IEEE.

Weber, M. (1961). Ethnic groups. Theories of society (ed. T. Parsons, E. Shils, K. Naegele, J. Pitts), 1, 305-9.

Yancey, A. K., Cole, B. L., Brown, R., Williams, J. D., Hillier, A., Kline, R. S., ... & McCarthy, W. J. (2009). A cross-sectional prevalence study of ethnically targeted and general audience outdoor obesity-related advertising. *Milbank Quarterly*, 87(1), 155-84.