International Journal of Information Management 51 (2020) 101934

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Emotional Text Mining: Customer profiling in brand management

Francesca Greco*, Alessandro Polli

Sapienza Università degli Studi di Roma, Italy

ARTICLE INFO

Keywords:
Emotional Text Mining
Brand management
Twitter
Network analysis
Customer profiling

ABSTRACT

The widespread use of the Internet and the constant increase in users of social media platforms has made a large amount of textual data available. This represents a valuable source of information about the changes in people's opinions and feelings. This paper presents the application of Emotional Text Mining (ETM) in the field of brand management. ETM is an unsupervised procedure aiming to profile social media users. It is based on a bottom-up approach to classify unstructured data for the identification of social media users' representations and sentiments about a topic. It is a fast and simple procedure to extract meaningful information from a large collection of texts. As customer profiling is relevant for brand management, we illustrate a business application of ETM on Twitter messages concerning a well-known sportswear brand in order to show the potential of this procedure, highlighting the characteristics of Twitter user communities in terms of product preferences, representations, and sentiments.

1. Introduction

The Internet's wide diffusion increases the opportunity for millions of people to surf the web daily to search and share information, ideas, interests, or any other forms of expression. Social media platforms, such as Twitter or Facebook, will increasingly play a crucial role in many areas as they enable direct, continuous and real-time communication (Chandler, Salvador, & Kim, 2018).

The most obvious outcome of such a communication process is a steady increase in user-generated content (He, Zha, & Li, 2013) about their activities, behaviors, attitudes, preferences and values, freely shared on such digital platforms; a circumstance that opens up great opportunities for research and marketing professionals, who can draw on this data repository cheaply and effectively.

The growing ease with which a professional can access a wide range of information on the markets in reference is a strategic resource for many business functions, including brand management (Shirdastian, Laroche, & Richard, 2017), which plays a crucial role in increasing the perceived value of a product, a product line or a brand over time and, ultimately, the brand equity.

Although the spread of social media has changed the tactics of brand management, the main purpose of branding remains to attract new consumers and their loyalty (Weber, 2009). Consumers often use the social media channel to express their feelings and opinions about products and consequently, their attitudes towards brands (Fan, Che, & Chen, 2017; Fronzetti Colladon, 2018). Not surprisingly, therefore, a crucial success factor for a brand management plan is to know what customers disclose when they share a text on social media (Jimenez-Marquez, Gonzalez-Carrasco, Lopez-Cuadrado, & Ruiz-Mezcua, 2019; Shiau, Dwivedi, & Lai, 2018).

The constant rise in the number of users on social media platforms make a large amount of data available that represents a relevant source of information. The scraping of social media platforms allows for the collection of huge amounts of textual data, typically unstructured, in a relatively short amount of time. Therefore, a methodology is needed to process unstructured data and to extract information. As shown by the literature, the online communication is analyzed by means of text mining procedures for different purposes, such as product planning (Jeong, Yoon, & Lee, 2017), marketing (AlAlwan, Rana, Dwivedi, & Algharabat, 2017; Kapoor et al., 2018), voting behavior forecasting (e.g., Greco, Maschietti, & Polli, 2017; Grover, Kar, Dwivedi, & Janssen, 2018), disaster management (Singh, Dwivedi, Rana, Kumar, & Kapoor, 2017), campaign surveys (Afful-Dadzie & Afful-Dadzie, 2017), and in the assessment of web sites, customer review effectiveness and customer perceptions of digital marketing (Antonacci, Fronzetti Colladon, Stefanini, & Gloor, 2017; Aswani, Kar, Ilavarasan, & Dwivedi, 2018; Gloor, Fronzetti Colladon, Giacomelli, Saran, & Grippa, 2017; Rekik, Kallel, Casillas, & Alimi, 2018; Singh, Irani et al., 2017). In addition, sentiment analysis is increasingly used in order to explore people's opinions and feelings (e.g., Aswani et al., 2018; Ceron, Curini, & Iacus, 2016; Gloor, 2017; Hopkins & King, 2010; Liu, 2012).

This paper aims to present a methodology for the analysis of

* Corresponding author at: Sapienza Università degli Studi di Roma, Via Capo d'Africa 37, 00184, Roma, Italy.
E-mail addresses: francesca.greco@uniroma1.it (F. Greco), alessandro.polli@uniroma1.it (A. Polli).

https://doi.org/10.1016/j.ijinfomgt.2019.04.007
Received 26 December 2018; Received in revised form 25 March 2019; Accepted 12 April 2019
Available online 14 June 2019

massive textual data, namely, Emotional Text Mining (ETM), and apply it in some of the typical areas of brand management, for example, brand identity management and brand loyalty monitoring. ETM is a particular kind of sentiment analysis based on a socio-constructivist approach and a psychodynamic model, which allows for the identification of the elements setting people's interactions, behavior, attitudes, expectations and communication. Thus, according to a semiotic approach to the analysis of textual data, ETM allows a social profiling to be performed. This has already been applied in different fields ranging from political debate, in order to profile social media users and to anticipate their political choices (Greco, Alaimo, & Celardo, 2018; Greco, Celardo, & Alaimo, 2018; Greco et al., 2017; Greco & Polli, 2019), to the professional training effectiveness at the Sapienza University of Rome (Cordella, Greco, Meoli, Palermo, & Grasso, 2018), to brain structure (Laricchiuta et al., 2018) and to the impact of the law on society (e.g., Greco, 2016; Cordella, Greco, Carlini, Greco, & Tambelli, 2018).

This paper is structured as follows; in Section 2, we present the theoretical approach; in Section 3, we present the ETM procedure; in Section 4, ETM is applied to a case study of a famous sportswear company following the launch of a new model of sports shoes, in order to extract useful information for business decision-making; in Section 5, we discuss the theoretical contribution of the main results, as well as the managerial implications; and in Section 6, we provide the conclusion.

2. A semiotic approach to sentiment analysis

Sentiment analysis is a field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities. It is also called opinion mining, since, frequently, the sentiment is considered a personal belief or judgment which is not founded on rationale reasoning, but on subjective emotion.

The use of a text mining approach to classify the sentiment of a text has been largely discussed in the literature, (e.g., Balbi, Misuraca, & Scepi, 2018; Bollen, Mao, & Zeng, 2011; Ceron, Curini, Iacus, & Porro, 2014; Fronzetti Colladon, 2018; Gloor, 2017; Jeong et al., 2017; Liu, 2012; Salvatore, Gennaro, Auletta, Tonti, & Nitti, 2012). Nevertheless, a text mining procedure has to refer, implicitly or explicitly, to a sociological or a psychological theoretical approach which explains the language production and the social interaction that sets the communication exchange. In order to be rigorous, a study should match the theoretical approach to the methodological one but, surprisingly, this aspect is apparently neglected by scholars (AlAlwan et al., 2017). Most of the literature on the text mining procedure draws particular attention to the methodology (word tagging, lexical structure of the sentence, statistical procedure, etc.), focusing on the manifest content of the text.

Most methods are based on a top-down approach where an a-priori coding procedure of terms, or text, is performed focusing on the manifest content of the word. Following a top-down approach, these methods use predefined content categories to semantically classify the text (e.g., Balbi et al., 2018; Liu, 2012). Each of these categories corresponds to a thematic dictionary containing all the words indicative of the content represented by that category.

Nevertheless, as highlighted by Saussure in a *Course in General Linguistics*, language is a system of signs that expresses a system of meaning. Even though the top-down approaches of text mining allow for a reliable and valid investigation, they present a major limitation, disregarding the contextual nature of the linguistic meaning (Carli & Paniccia, 2002; Salvatore & Freda, 2011). Therefore, a term can assume a specific meaning according to its association to the other terms in the text.

As stated by Liu (2012), it is not sufficient to classify the sentiment lexicon in order to perform a sentiment analysis because a term, classified as a positive or negative sentiment word, may have an opposite orientation depending on the context. In fact, the meaning of a word is polysemic and is subject to the way it combines with other words in a communicative interaction, i.e., it depends on its association with other words. For example, "bomb" usually indicates a negative sentiment, e.g., "There was another truck bomb explosion this morning at the market in Sadr City", but it can also imply a positive sentiment of admiration, e.g., "she's a sex bomb!". Therefore, the presence, or absence, of sentiment words in a sentence does not necessarily imply the possibility of classifying a sentiment. That is to say, a sentence containing sentiment terms may be neutral, which happens frequently in questions or conditional sentences, and a sentence without a sentiment word may express an opinion. Moreover, sarcastic sentences, with or without sentiment words, are difficult to classify (Liu, 2012).

Based on these considerations, Emotional Text Mining (ETM) (Greco, 2016) is a text mining procedure that, by means of its bottom-up logic, allows for a context-sensitive text mining approach on unstructured data, which constitutes 95% of big data (Gandomi & Haider, 2015). ETM is an unsupervised text mining procedure, based on a socio-constructivist approach and a psychodynamic model. According to this approach, sentiment is not only the expression of a mood, but also the evidence of a latent and social thinking process that sets people interactions, behavior, attitudes, expectations and communication.

We know that a person's behavior depends not only on their rationale thinking but also, and sometimes most of all, on their emotional and social way of mental functioning (Carli, 1990; Moscovici, 2005; Salvatore & Freda, 2011). In other words, people consciously categorize reality and, at the same time, unconsciously symbolize it emotionally, in order to adapt to their social environment (Fornari, 1976). The conscious categorization and unconscious symbolization are two parallel mental processes that follow two different functioning rules, i.e. two logic (Matte Blanco, 1975). The unconscious symbolization is social, as people generate it interactively and share the same emotional meanings through this interaction (Greco, 2016). Since communication and behavior are the outcome of this social mental functioning, it is possible to analyze the communication (text) to infer the social mental functioning (symbolic matrix) and explain, or forecast, people's behavior in different contexts. Moreover, explaining or forecasting people's behavior by means of their social media communication is relevant for business management (Gloor, 2017; He et al., 2013; Lipizzi, Iandoli, & Ramirez Marquez, 2015; Liu, 2012).

Due to the fact that the conscious process sets the manifest content of the communication, i.e. what is communicated, the unconscious process can be inferred through how it is communicated, namely, the words chosen to communicate and their association within the text. We consider that people emotionally symbolize an event, or an object, and socially share this symbolization. The words they choose to discuss an event, or object, is the product of the socially-shared, unconscious mental functioning (Greco, 2016).

3. The Emotional Text Mining procedure

ETM is an unsupervised text mining procedure allowing for the detection of the symbolic matrix and the representation and the sentiment of an entity, e.g. a specific brand. These three elements are interconnected, as the symbolic matrix generates the representation (Carli, 1990) and the representation sets the sentiment as well as behavior (Moscovici, 2005). Moreover, they imply different levels of generalization and awareness. While a person is aware of his/her sentiment, she/he is not directly aware of the representation (Moscovici, 2005), nor is she/he aware of the symbolic matrix, which is unconscious and socially shared. For this reason, the ETM procedure allows for the detection of both the semantic and the semiotic aspects conveyed by the communication.

While the mental functioning proceeds from the semiotic level to the semantic one in generating the text, the statistical procedure simulates the inverse process of the mental functioning, from the semantic level to the semiotic one. For this reason, ETM performs a sequence of synthesis procedures, from the reduction of the type to

lemma and the selection of the keywords to the clustering and the factorial analysis, in order to identify the semiotic level (the symbolic matrix), starting from the semantic one (the word co-occurrence) (Cordella, Greco, & Raso, 2014; Greco et al., 2017).

In order to detect the associative links between the words and to infer the symbolic matrix determining their coexistence into the text, first we perform a bisecting k-means algorithm (Savaresi & Boley, 2004; Steinbach, Karypis, & Kumar, 2000), limited in the number of partitions, excluding all the text that does not have at least two keywords co-occurrence to classify the text. We have selected this clustering procedure as it is the most commonly used one in the semiotic approach (Greco, 2016). As in the literature, the identification of a reliable methodology for results evaluation is still controversial (e.g., Misuraca, Spano, & Balbi, 2018) and three clustering validation measures are taken into account in order to identify the optimal solution: the Calinski-Harabasz, the Davies-Bouldin and the intraclass correlation coefficient (ICC) indices.

Next, we perform a correspondence analysis (Lebart & Salem, 1994) on the cluster per keywords matrix. While the cluster analysis allows for the detection of the representations, the correspondence analysis detects the symbolic matrix.

The interpretation process proceeds from the highest level of synthesis to the lowest one, simulating once again the mental functioning. Therefore, first we interpret the factorial space according to word polarization (Greco, 2016), in order to identify the symbolic matrix setting the communication. Then, we interpret the cluster according to their location in the factorial space and to the words characterizing the context units classified in the cluster, in order to identify the representation. Finally, the sentiment is defined in relation to the elements characterizing the representations (positive, neutral, or negative), and it is calculated according to the number of messages classified in the cluster.

4. A case study

4.1. Research questions

In the ETM, we perform a cluster analysis on the term per document matrix in order to detect word co-occurrence, considering each document as a vector of n dimensions (n = number of keywords). Alternatively, we can consider the vector space resulting from terms and documents as an adjacency matrix, in order to extract relationship patterns between terms and documents, considering the corpus as a network of texts and words (Iezzi, 2012).

This study collected the tweets containing the name of a well-known sportswear brand in a specific period of time, and applied the ETM and a network analysis (NA), based on a community detection algorithm in order to analyze unstructured text content with a bottom-up approach. Specifically, the study attempts to answer the following questions:

- What are the general categories organizing the communication about the brand?
- What representations of the brand can be found in the Twitter messages?
- What are the sentiments towards the brand?
- What are the main differences in pattern detection among the two text mining procedures (EMT vs NA)?

4.2. Methodology

In order to explore the representations of the sportswear brand in Twitter communications, we scraped all the messages from the Twitter repository, written in English, containing the sportswear brand name from November 29th to December 3rd, 2018. The data extraction was carried out with the *twitteR* package of R Statistics (Gentry, 2016). A sample of 107,500 tweets was made up of 63.7% of retweets and resulted in a large size corpus. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the percentage of hapax (Giuliano & La Rocca, 2010).

First, the data were cleaned and pre-processed with the software T-Lab (Lancia, version T-Lab Plus 2018) and keywords were selected. In particular, we used lemmas as keywords instead of type, filtering out the lemma of the sportswear brand and those of low rank of frequency (Bolasco, 1999; Greco, 2016). Then, on the tweets per keyword matrix, we performed a cluster analysis with a bisecting k-means algorithm based on cosine similarity (Savaresi & Boley, 2004) limited to 20 partitions, excluding all the tweets that did not have at least two keywords co-occurrence. In order to choose the optimal solution, we calculated the Calinski-Harabasz, the Davies-Bouldin and the intraclass correlation coefficient (ρ) indices.

Then, we performed a correspondence analysis (Lebart & Salem, 1994) on the cluster per keywords matrix, and the sentiment was calculated according to the number of messages classified in the cluster and its interpretation. Finally, we performed a network analysis with a community detection model, the Louvain's algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). We chose this method as it is suitable for a large network of textual data.

4.3. Findings

After the release of a new model of shoes on November 16th, 2018, the number of messages produced from November 29th to December 3rd were, on average, more than 20,000 tweets per day. The corpus preprocessing determined a loss of 10% of the messages (n = 96,361) resulting in a large size corpus of 1,313,025 tokens. On the basis of the large size of the corpus, both lexical indicators highlight its richness (TTR = 0.02; Hapax percentage = 45.0) and indicate the possibility of proceeding with the statistical analysis, which was performed with the 758 keywords selected.

4.3.1. Emotional Text Mining results

The results of the cluster analysis show that the keywords selected allow for the classification of 92.2% of the tweets. The clustering validation measures show that the optimal solution is five clusters (Calinski-Harabasz = 2763.3; Davies-Bouldin = 1.61; ρ = 0.111).

The correspondence analysis detected six latent dimensions, and the explained inertia for each factor is reported in Table 1. In Fig. 1, we can appreciate the factorial space of the sportswear brand emerging from the English tweets. It shows how the clusters are placed in the factorial space produced by the first three factors, explaining 82.2% of the inertia.

As shown in Table 2, ultimately, the Twitter users symbolize the brand by means of four main categories: the type of brand, the buying preferences, the use of the sportswear and the model. The first factor distinguishes the customer's reason for choosing the brand, because it is *cool* (fashionable) or because it produces sportswear; the second factor refers to the buyer's preferences, the customer who likes to look for a bargain and the regular customer. The third factor reflects the use of the sportswear, differentiating between leisure and competition. Finally, the fourth factor concerns the customer's preferences for new models,

Table 1
Correspondence analysis results.

Factor	Eigenvalue	%	Cumul. %	Label	Negative Pol.	Positive Pol.
1	0.674	33.74	33.74	Brand	Fashion	Sport
2	0.542	27.15	60.89	Customer	Bargain Hunter	Regular
3	0.516	25.81	86.69	Use	Day Time	Pro
4	0.266	13.31	100.00	Model	Remake	News

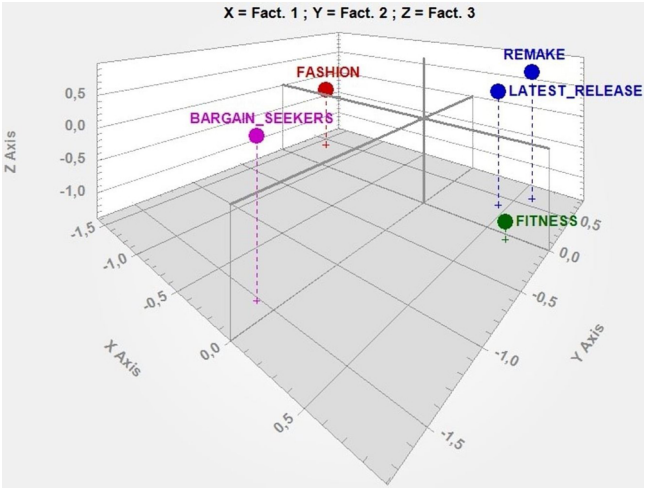

Fig. 1. Factorial space set by the first three factors.

Table 2
Cluster location in the symbolic space.

Cluster	Label	Factor 1 Brand	Factor 2 Custom	Factor 3 Use	Factor 4 Model
1	Remake	Sport	Regular	Pro	Remake
2	Bargain Hunters	Fashion	Bargain Hunter	Pro	
3	Fitness	Sport		Day Time	
4	Fashion	Fashion	Regular	Day Time	
5	Latest Release	Sport	Regular	Pro	New

distinguishing between the customer who prefers the remake of an outdated success and the one who prefers the new innovation.

The interpretation of the factorial space highlights the symbolic categories by which people, in general, emotionally categorize the brand, and support the cluster interpretation according to their location in the symbolic space (Table 2).

The five clusters are of different sizes (Table 3) and reflect different brand representations. In the first cluster, the brand is perceived as a company able to produce good quality sportswear used by famous sport champions, whose customers appreciate over time; in the second cluster, the brand is considered as a valuable object that can be collected or exchanged by bargain hunters; in the third cluster, the brand is represented as sportswear useful for leisure activities by sport lovers, i.e. people who like to be fit; in the fourth cluster, the brand is represented as the producer of fashion sportswear. The customers seem to be more interested in the design rather than in the technology, as there are words like *icon, good, fashion* and *lovely*. Finally, in cluster five, the brand is perceived as a trustworthy sportswear company, as in cluster one, but customers seem to be more interested in the most recent model.

It is interesting to note that each cluster is frequently associated with a specific color and sportswear model. For example, the first cluster is associated with the model *airmax* and the color *black, white* (Table 3), *red* (f = 1355) and many others (*blue, grey, gold, silver, orange, green, yellow* and *brown*) appearing in a small number of messages, ranging from 741 to 218 tweets. Moreover, it seems to be connected to gender as the term *man* appears in 1625 tweets. Only in the bargain hunters' cluster are there neither model nor color, and where words probably connected to an evaluation appear (*fuck, hate*).

From the interpretation of the clusters, we detected five different representations of the brand connected to a specific community of Twitter users who seem to share a similar approach to the brand. As all the representations seem to be mainly positive, we grouped the representations in two sentiments: sport lovers and fashion lovers (Fig. 2).

Table 3
Brand representations and sentiment.

Cluster	Tot Tweet classified	Size	Label	keyword	N Tweet	Sentiment
1	21,473	24.2	Milestone	Air	16,470	Love Sportswear
				Max	7,213	
				size	4,634	
				Jordan	4,390	
				black	3,898	
				force	2,865	
				white	2,708	
				Retro	2,060	
2	17,313	19.5	Bargain Hunters	shoe	4,143	Love Fashion
				buy	2,725	
				give away	2,407	
				thank	2,275	
				enter	2,241	
				group	2,078	
				bot	2,013	
				AMNotify	1,968	
3	17,279	19.5	Fitness	white	5,499	Love Sportswear
				pair	4,588	
				zoom	4,451	
				time	4,240	
				day	4,238	
				Fly	3,815	
				phone	3,365	
				playstation	3,207	
4	24,051	27.1	Fashion	back	6,514	Love Fashion
				icon	6,199	
				tee	5,812	
				wear	3,470	
				good	2,608	
				share	2,019	
				check out	1,711	
				sock	1,681	
5	8,702	9.8	Latest release	free	2,002	Love Sportswear
				ship	1,982	
				gt	1,366	
				react	1,305	
				low	1,125	
				drop	1,062	
				Kyrie	1,017	
				dunk	974	

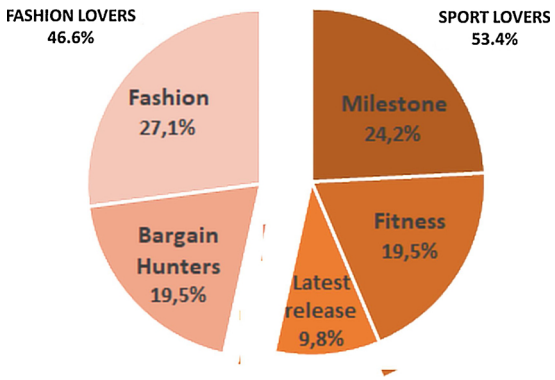

Fig. 2. Sentiment on the sportswear brand.

We classified as sport lovers, people who love the milestone model, those who love to be fit and those who like new releases, as they seem to be mostly focused on the technological innovation and its use. On the other hand, we considered the bargain hunters and the fashion

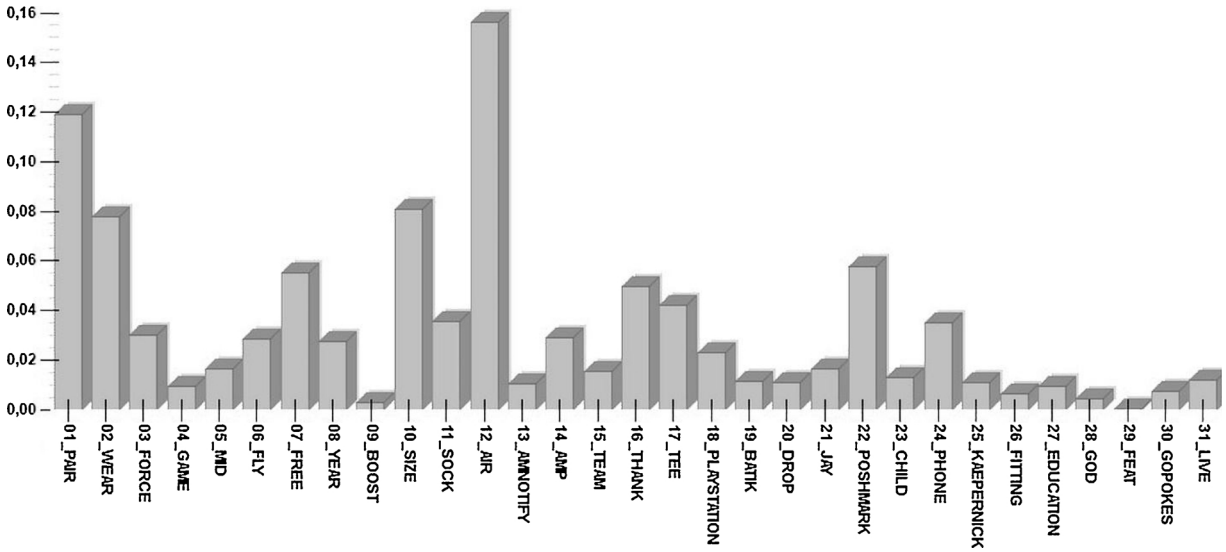

Fig. 3. Size of the communities.

customers as fashion lovers because they are more focused on the brand's image.

4.3.2. Community detection model with the Louvain's algorithm

The community detection algorithm identified 31 communities, the size of each is shown in Fig. 3. There are a large number of small communities and few larger ones. Among the largest five (community n. 01, 02, 10, 12, 22), the first community "Pair" and twelfth community "Air" are similar to cluster 1 (Milestones) and cluster 2 (Bargain Hunters) of the ETM in words composition, while the other three communities do not share a common lexical profile with the EMT clusters.

The relationship within the 758 keywords is shown in Fig. 4. Due to the large number of terms, the interpretation of the graph is relatively challenging. Comparing the two text mining procedures, the ETM and the NA, it seems that the first one identifies a smaller number of partitions, thus being more effective in the identification of the Twitter user's lexical profiles. Quite possibly, a top-down approach might be more effective while using the NA, as it could reduce the sparseness of the matrix.

5. Discussion

This paper presents the application of Emotional Text Mining in the field of brand management (e.g., Fronzetti Colladon, 2018), with particular emphasis on the themes of brand identity management and loyalty brand monitoring. The semiotic approach of the ETM allows for the profiling of the customers of a well-known sportswear brand by profiling Twitter users. In other words, we were able to identify Twitter users' symbolic categories and representations of the sportswear brand, and to measure their sentiments. The case study was used as an example to illustrate the potentiality of ETM in the field of brand management, but its application can easily be extended depending on the analyst's interests.

5.1. Theoretical contributions

Although our case study was limited to Twitter, which may not allow for the results to be generalized with regard to other platforms (Kapoor et al., 2018), ETM can be applied to a variety of languages and documents, from social media and media documents (e.g., Greco, 2016; Greco et al., 2017) to interviews or focus groups (e.g., Cordella, Greco, Meoli et al., 2018). Moreover, ETM applies a bottom-up approach to unstructured data, which constitutes 95% of big data (Gandomi & Haider, 2015), and could be usefully applied to this volume of data for real-time analytics, owing to the fact that the analyst's intervention is only required for the interpretation of the output. For this reason, we think that ETM is likely to become a useful research tool due to the growth in the use of social media, and the usefulness of data analytics aimed at supporting businesses in converting large volumes of messages into meaningful information, thereby supporting decision-making (Gandomi & Haider, 2015).

Unlike the sentiment analysis based on a supervised procedure, e.g. machine learning (Ceron et al., 2016; Hopkins & King, 2010), in which the researcher's interpretation is performed at the beginning of the analysis in order to build the training set, in ETM the interpretation is performed at the end of the statistical analysis. The advantage of the ETM approach is to identify the elements connected with a specific sentiment, as the representations are a system of values, ideas, and practices setting people's interaction and behavior.

Due to the limited number of characters in a tweet and to its lexical peculiarity, ETM could be less accurate in the classification of messages, as it is based on a word co-occurrence logic. Nevertheless, we have addressed this problem by adopting a specific keyword selection criteria (Greco et al., 2017) that allows for classifying practically all the messages.

The application of ETM is interesting, as it complements the results of market research (e.g., Dwivedi, Kapoor, & Chen, 2015; Gloor, 2017; He et al., 2013; Liu, 2012) and focuses on groups through virtually continuous monitoring of brand perception by potential customers. The advantage of applying this methodology is the extraction of structured information, which is therefore highly significant, from an unstructured collection of texts from a potentially huge quantity of data.

The application of ETM allowed us to identify four symbolic categories that set the communication about the brand: the value of the brand, the type of customer, the customer's use of the product and the preferences revealed by the consumer. Within these symbolic categories, ETM detects five brand representations and the characteristics pertaining to each community of customers, regarding their product preferences (model, color, purchase choices, etc.) and their brand sentiment (fashion lovers or sport lovers). With regard to the use of network analysis, it is interesting to note that this methodology seems to be more appropriate for the analysis of content. However, network analysis identifies a large number of communities, which is less effective in reducing the complexity of textual data. The use of a top-down approach with a multi-stage agglomeration strategy (Balbi et al., 2018)

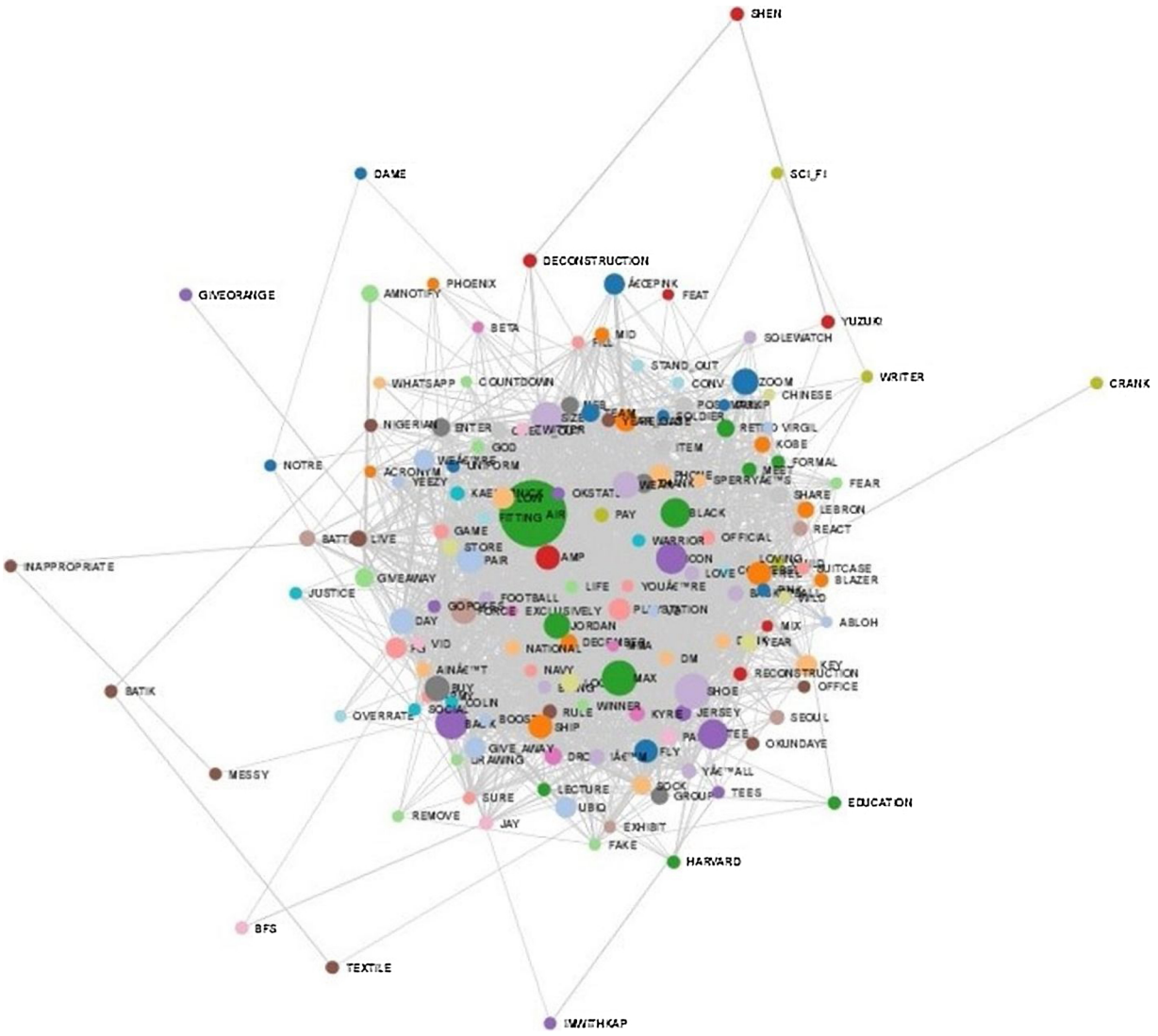

Fig. 4. Network of the keywords.

could solve this problem.

5.2. Implication for practice

In addition to these considerations, being essentially of a theoretical nature, there are some practical reasons that make the methodology discussed previously particularly interesting in view of its operational implications.

Firstly, text mining has proved to be a valuable tool in business intelligence and in social media marketing (Dwivedi et al., 2015; Lin, Li, & Wang, 2017; Xu, Wang, Li, & Haghighi, 2017). ETM is a fast, cheap and simple way to extract more meaningful information from large collections of texts. Indeed, the application of ETM greatly reduces the complexity of textual data, while preserving its information content. Such reduction is performed through the classification of texts and the identification of factors that explain the diversity between the different clusters. The use of an unsupervised procedure allows for obtaining results without any intervention, as there is no need to train a classification algorithm. The only intervention required is the final interpretation of the results by an operator.

The second practical reason is that reading and interpreting the results becomes relatively straightforward. As we have clarified above, considering that the scraping of Twitter or any other social media can lead to the collection of hundreds of thousands of texts, it appears important to extract from this large amount of textual data only the essential information, which in the case of ETM is easy to achieve even for the average user. After becoming familiar with these tools, the social media specialist could drastically reduce the time spent in reading and classifying the textual data (often done manually, an activity which could lead to misclassification) and focus on the more creative stages of his/her work. For the same reason, ETM can easily be implemented by small firms (Braojos-Gomez, Benitez-Amado, & Llorens-Montes, 2015) that are willing to develop a social media competence.

Finally, the radical simplification of the textual data preprocessing step opens up the possibility of repeating the surveys frequently and with little effort, making the monitoring of target markets on social media a virtually continuous activity, carried out in real-time. This aspect is very important, as it makes the social media specialist more responsive to detecting the rise of new trends, aspirations, and needs.

To summarize, the introduction of ETM in the social media manager's task list can provide clear advantages in terms of cost reduction, limiting the most time-consuming activities and, ultimately, increasing productivity and effectiveness identifying customers' profiles and social media communities.

6. Conclusion

The widespread use of the Internet and the constant increase in

users of social media platforms has made a large amount of textual data available. This represents a valuable source of information regarding changes in the opinions and feelings of people, with reference to the most disparate topics. The extraction of textual data from a social media platform allows for the collection of a large amount of data, typically unstructured, in a reasonably short time. It is, therefore, necessary to apply technologies and methods of analysis to big data, aimed at extrapolating from this mass of textual data information, and ultimately, knowledge, useful for businesses and their brand managers.

After a brief survey of the literature, we presented the results of an ETM applied to a typical problem of brand management, related to the management of brand identity and the monitoring of the brand loyalty. The results obtained make it possible to identify the area in which this method provides the best results, as well as highlight the main limitations. More specifically, ETM seems to be more effective on large collections of textual data, when the aim is to identify communities of potential customers, with reference both to their perception of brand value and brand loyalty.

Hence, ETM seems to be applicable in the field of brand management, as it allows for the identification of groups of customers who, according to their lexical profile, share the same brand representation. Moreover, this method allows for the identification of the general categories which can be used to organize communication about the brand on social media. Even though it is a case study, our research could easily be enriched, combining the structured information obtained by applying ETM with the profiling data of Twitter users. This, then, allows for the definition of profiles, which correspond to specific customer segments, with significant advantages in terms of costs and timeliness for obtaining results.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Afful-Dadzie, E., & Afful-Dadzie, A. (2017). Liberation of public data: Exploring central themes in open government data and freedom of information research. *International Journal of Information Management, 37*(6), 664–672.

AlAlwan, A., Rana, N. P., Dwivedi, Y. K., & Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics, 34*(7), 1177–1190.

Antonacci, G., Fronzetti Colladon, A., Stefanini, A., & Gloor, P. (2017). It is rotating leaders who build the swarm: Social network determinants of growth for healthcare virtual communities of practice. *Journal of Knowledge Management, 21*(5), 1218–1239.

Aswani, R., Kar, A. K., Ilavarasan, P. V., & Dwivedi, Y. K. (2018). Search engine marketing is not all gold: Insights from twitter and SEOClerks. *International Journal of Information Management, 38*(1), 107–116.

Balbi, S., Misuraca, M., & Scepi, G. (2018). Combining different evaluation systems on social media for measuring user satisfaction. *Information Processing & Management, 54*(4), 674–685.

Blondel, V. D., Guillaume, J. G., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment, 10*, 1–12.

Bolasco, S. (1999). *Analisi multidimensionale dei dati: metodi, strategie e criteri d'interpretazione.* Roma, IT: Carocci.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Braojos-Gomez, J., Benitez-Amado, J., & Llorens-Montes, F. J. (2015). How do small firms learn to develop a social media competence? *International Journal of Information Management, 35*(4), 443–458.

Carli, R. (1990). Il processo di collusione nelle rappresentazioni sociali. *Rivista di Psicologia Clinica, 3*, 282–296.

Carli, R., & Paniccia, R. M. (2002). *L'Analisi Emozionale del Testo: Uno strumento psicologico per leggere testi e discorsi.* Milano, IT: Franco Angeli.

Ceron, A., Curini, L., & Iacus, S. M. (2016). ISA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences, 367–368*, 105–124.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16*(2), 340–358.

Chandler, J. D., Salvador, R., & Kim, Y. (2018). Language, brand and speech acts on Twitter. *Journal of Product and Brand Management, 27*(4), 375–384.

Cordella, B., Greco, F., & Raso, A. (2014). Lavorare con Corpus di Piccole Dimensioni in Psicologia Clinica: Una Proposta per la Preparazione e l'Analisi dei Dati. In E. Nee, M. Daube, M. Valette, & S. Fleury (Eds.). *Actes JADT 2014, 12es Journées internationales d'Analyse Statistque des Données Textuelles* (pp. 173–184). Paris, FR: JADT.org.

Cordella, B., Greco, F., Carlini, K., Greco, A., & Tambelli, R. (2018). Infertilità e procreazione assistita: evoluzione legislativa e culturale in Italia. *Rassegna di Psicologia, 35*(3), 45–56. https://doi.org/10.4458/1415-04.

Cordella, B., Greco, F., Meoli, P., Palermo, V., & Grasso, M. (2018). Is the educational culture in Italian Universities effective? A case study. In D. F. Iezzi, L. Celardo, & M. Misuraca (Eds.). *JADT' 18: Proceedings of the 14th International Conference on Statistical Analysis of Textual Data* (pp. 157–164). Rome, IT: Universitalia.

Dwivedi, Y. K., Kapoor, K. K., & Chen, H. (2015). Social media marketing and advertising. *The Marketing Review, 15*, 289–309.

Fan, Z. P., Che, Y. J., & Chen, Z. Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of Business Research, 74*, 90–100.

Fornari, F. (1976). *Simbolo e codice: Dal processo psicoanalitico all'analisi istituzionale.* Milano, IT: Feltrinelli.

Fronzetti Colladon, A. (2018). The Semantic Brand Score. *Journal of Business Research, 88*, 150–160.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Gentry, J. (2016). *R based Twitter client. R package version 1.1.9.*

Giuliano, L., & La Rocca, G. (2010). *Analisi automatica e semi-automatica dei dati testuali, Vol. II.* Milano: Led.

Gloor, P. A. (2017). *Sociometrics and human relationships: Analyzing social networks to manage brands, predict trends, and improve organizational performance.* London, UK: Emerald Publishing Limited.

Gloor, P., Fronzetti Colladon, A., Giacomelli, G., Saran, T., & Grippa, F. (2017). The impact of virtual mirroring on customer satisfaction. *Journal of Business Research, 75*, 67–76.

Greco, F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale.* Milano, IT: Franco Angeli.

Greco, F., & Polli, A. (2019). Vaccines in Italy: The Emotional Text Mining of social media. *Rivista Italiana di Economia Demografia e Statistica, 73*(1), 89–98.

Greco, F., Maschietti, D., & Polli, A. (2017). Emotional Text Mining of social networks: The French pre-electoral sentiment on migration. *Rivista Italiana di Economia Demografia e Statistica, 71*(2), 125–136.

Greco, F., Alaimo, L., & Celardo, L. (2018). Brexit and Twitter: The voice of people. In D. F. Iezzi, L. Celardo, & M. Misuraca (Eds.). *JADT' 18: Proceedings of the 14th International Conference on Statistical Analysis of Textual Data* (pp. 327–334). Rome, IT: Universitalia.

Greco, F., Celardo, L., & Alaimo, L. M. (2018). Brexit in Italy: Text mining of social media. In A. Abbruzzo, D. Piacentino, M. Chiodi, & E. Brentari (Eds.). *Book of short papers SIS 2018* (pp. 767–772). Milano: Pearson.

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2018). *Polarization and acculturation in US Election 2016 outcomes–can twitter analytics predict changes in voting preferences.* Technological Forecasting and Social Changehttps://doi.org/10.1016/j.techfore.2018.09.009.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33*(3), 464–472.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229–247.

Iezzi, F. D. (2012). Centrality measures for text clustering. *Communications in Statistics – Theory and Methods, 41*(16–17), 3179–3197.

Jeong, B., Yoon, J., & Lee, J. M. (2017). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management.* https://doi.org/10.1016/j.ijinfomgt.2017.09.009.

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management, 44*, 1–12.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers, 20*(3), 531–558.

Lancia, F. (2018). *User's manual: Tools for text analysis. T-Lab version Plus 2018.*

Laricchiuta, D., Greco, F., Piras, F., Cordella, B., Cutuli, D., Picerni, E., Assogna, F., Lai, C., Spalletta, G., & Petrosini, L. (2018). "The grief that doesn't speak": Text mining and brain structure. In D. F. Iezzi, L. Celardo, & M. Misuraca (Eds.). *JADT' 18: Proceedings of the 14th International Conference on Statistical Analysis of Textual Data* (pp. 419–427). Rome, IT: Universitalia.

Lebart, L., & Salem, A. (1994). *Statistique textuelle.* Paris, FR: Dunod.

Lin, X., Li, Y., & Wang, X. (2017). Social commerce research: Definition, research themes and the trends. *International Journal of Information Management, 37*, 190–201.

Lipizzi, C., Iandoli, L., & Ramirez Marquez, J. E. (2015). Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using twitter streams. *International Journal of Information Management, 35*(4), 490–503.

Liu, B. (2012). *Sentiment analysis: Mining opinions, sentiments, and emotions. Sentiment analysis: Mining opinions, sentiments, and emotions.* Morgan & Claypool1–367.

Matte Blanco, I. (1975). *The unconscious as infinite sets: An essay in bi-logic.* London, UK: Duckworth.

Misuraca, M., Spano, M., & Balbi, S. (2018). BMS: An improved Dunn index for document clustering validation. *Communications in Statistics: Theory and Methods,* 1–14.

Moscovici, S. (2005). *Le rappresentazioni sociali.* Bologna, IT: Il Mulino.

Rekik, R., Kallel, I., Casillas, J., & Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. *International Journal of Information Management, 38*, 201–216.
Salvatore, S., & Freda, M. F. (2011). Affect, unconscious and sensemaking. A psychodynamic, semiotic and dialogic model. *New Ideas in Psychology, 29*(2), 119–135.
Salvatore, S., Gennaro, A., Auletta, A. F., Tonti, M., & Nitti, M. (2012). Automated method of content analysis: A device for psychotherapy process research. *Psychotherapy Research, 22*(3), 256–273.
Savaresi, S. M., & Boley, D. L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis, 8*(4), 345–362.
Shiau, W.-L., Dwivedi, Y. K., & Lai, H.-H. (2018). Examining the core knowledge on Facebook. *International Journal of Information Management, 43*, 52–63.
Shirdastian, H., Laroche, M., & Richard, M. O. (2017). Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. *International Journal of Information Management.* https://doi.org/10.1016/j.ijinfomgt.2017.09.007.
Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2017). Event classification and location prediction from tweets during disasters. *Annals of Operations Research,* 1–21.
Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research, 70*, 346–355.
Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *KDD workshop on text mining, vol. 400*, 525–526.
Weber, L. (2009). *Marketing to the social web: How digital customer communities build your business.* London: Wiley.
Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of Information Management, 37*, 673–683.

Francesca Greco received her PhD in Sociology at the Sapienza University of Rome and her PhD in Psychology at the University of Paris Descartes. She is currently the Research Manager of Prisma S.r.l., and she is qualified as Associate Professor in General Sociology. She is assistant professor in "Quantitative models for socio-economic analysis" at the Sapienza University of Rome and she is a member of the Italian Sociological Association and of the Italian Statistical Society. She is an expert in textual analysis and has developed a text mining procedure to perform social profiling. Her areas of interest are focused on psychosocial processes in the field of health care, disability, organizational management, political debate and deviance.

Alessandro Polli took a PhD in economic analysis of social phenomena at Sapienza University of Rome, where actually he teaches economic statistics and quantitative methods for economics. He has been scientific advisor of several public institutions, like Bank of Italy, and the Italian Presidency of the Council of Ministers. He is a member of the Italian Society of Economic, Demography, and Statistics. His research fields are statistic methods for market research, sustainable development and quality of life, assessment of the economic impacts of the migrations, assessment of the economic impacts of the new technologies on the job market.