A network-based concept extraction for managing customer requests in a social media care context

Michelangelo Misuraca[a,*], Germana Scepi[b], Maria Spano[b]

[a] *Department of Business and Law, University of Calabria, Italy*
[b] *Department of Economics and Statistics, University of Naples Federico II, Italy*

Abstract

Web 2.0 changed everyday life in many aspects, including the whole system that orbits around the purchase of products and services. This revolution necessarily involved also companies, because customers became increasingly demanding. The diffusion of social media platforms pushed customers to prefer this channel for quickly obtaining information and feedback about what they want to buy, as well as for asking help after the selling. In this framework, many organisations adopted a new way of providing assistance known as social customer care. A direct link to companies allows customers to obtain real-time solutions. In this paper, we introduce a new strategy for automatically managing the information listed in the requests that customers send to the social media accounts of companies. Our proposal relies on the use of network techniques for extracting high-level structures from texts, highlighting the different concepts expressed into the customers' written requests. The texts can be then organised on the basis of this new emerging information. An application to the requests sent to the AppleSupport service on Twitter shows the effectiveness of the strategy.

Keywords: customer care, textual data, network analysis, community detection

1. Introduction

More and more, the value of a product (or a service) encompasses both characteristics and utilities of the good itself as well as what the companies offer in terms of additional services. Customer care is then at the heart of all successful companies because improving the relationships with customers is one of the keys for enhancing their loyalty (Jiang et al., 2016). The traditional definition of customer care describes this task as taking care of customers to ensure their satisfaction after-sales (Grönroos, 1978), managing reports and complaints in the best ways and in the best times. In the modern era customers became

[*] Corresponding author. Tel.: +39 0984 49 2450
Email address: michelangelo.misuraca@unical.it (Michelangelo Misuraca)

increasingly demanding, making the challenge for the companies harder (Newell, 2001). The incredible progress of computer technology and the growth of the Internet has hastened the transition from analogue to digital data communication. This revolution necessarily involved both customers and companies. Because of the Web, everyday life changed in many aspects, also influencing the whole system that orbits around the purchase of a product or a service. The need for a change of direction in customer service increased and many organisations already adopted a new way of providing assistance via social media platforms, in the framework of the so-called *social customer care*.

Social media represent a new way for exploring several domains of everyday life (e.g. Kaplan & Haenlein, 2010; Kapoor et al., 2018; Shiau et al., 2018). With the development of the Web 2.0, sharing personal feelings and judgments with others has become a common habit. Users prefer more and more social media also to obtain information and feedbacks concerning what they want to buy (Alalwan et al., 2017), and to ask assistance and rapid solutions to their problems after the selling, interacting with each other (Rosenbaum & Massiah, 2007) and directly communicating with the companies (Heller Baird & Parasnis, 2011). This task is crucial for any business since a dissatisfied customer is a potential danger to the companies. Not helping a customer in time or not helping him at all can trigger a vicious circle of bad reputation (Shirdastian et al., 2017), destined to become viral (Karakaya & Ganim Barnes, 2010; Laroche et al., 2013), stated that electronic word-of-mouth can influence customers' behaviour (Sandes & Urdan, 2013). Therefore, it is strategic for companies to analyse the huge amount of texts produced on social media (Jimenez-Marquez et al., 2019), in a knowledge discovery and management perspective (Stieglitz et al., 2018).

Texts can express a wide and rich range of information, encoding this information in a form difficult to process from a quantitative viewpoint. In Text mining, the most common algebraic model for representing documents is the *vector space model* (Salton et al., 1975): a document is seen as a vector in the (extremely sparse) space spanned by the terms. Because of this model, each document contains a lot of noise. Documents are seen as *bag-of-words*, i.e. as an unordered set of terms, disregarding grammatical and syntactical roles. An effective text analysis necessarily requires a reduction of the original space dimensionality, because only a part of the collected data is relevant and informative for the phenomena of interest. It is possible to consider either feature selection and feature extraction techniques. Feature selection allows filtering a subset of the original terms, by excluding the less informative and discriminative ones or considering only the most relevant ones, with respect to a given criterion. Feature extraction performs a reduction of the original space by combining the terms into new entities. One of the main differences is that selection techniques retain the original meaning of terms, where extraction techniques require an additional effort in interpreting the results.

Social media push customers to produce short-length texts and use a colloquial register. This means that it is not easy to implement automatic systems or analytics (Pinto et al., 2010). Short text collections include documents containing a few terms. The main characteristic of this kind of texts is that the frequency of the terms is relatively low in comparison with their frequency in long documents. At the same time, the average document similarity of short text collections is very low, making difficult grouping texts related to similar topics. Here we propose a novel strategy designed for managing texts like posts and comments shared onto social media when we have a few significant terms for characterising the different groups of texts. The basic idea is that textual data can be processed at different levels, e.g. we can consider single terms or subsets of terms identifying different concepts. Concepts are identified by using tools designed for network analysis. Differently, from the bag-of-words, a network approach allows visualising the relations among the terms of a document collection by recovering the context of use of the terms themselves. It is possible then to categorise the documents with respect to the different concepts they list and interpret each group of documents with respect to the terms they contain. In a managerial perspective, this strategy can help companies to highlight the main complaints and needs of the customers, addressing the requests to the section of the technical or administrative support that more suitably can offer assistance.

The paper is structured as follows. In Section 2, reference literature is reviewed. Section 3 introduces the problem and describes the proposed strategy. In Section 4, the effectiveness of the proposal is showed by analysing a set of tweets sent to the AppleSupport account on Twitter. Section 5 discusses theoretical and practical implications. Final remarks and some possible future development of the research are considered in Section 6.

2. Background and theoretical framework

Since Web 2.0 became widely popular, users started producing new texts daily – in the form of *user-generated contents* (UGC) – and publishing them as blog posts, online reviews or opinions in several generalist or specialised social media. In this context, short-length texts are the prevalent way of creating and sharing information. This kind of texts is usually noisier and less topic-focused, with respect to *standard* documents, since they consist of a few terms (Yan et al., 2009). Moreover, they are characterised by insufficient context information and a peculiar grammatical style. Texts published on social media are often barely sentence-like fragments, not following lexical and syntactic conventions assumed by many language processing tools. Textual analyses usually follow statistical reasoning, aiming at finding trends, revealing patterns or explaining relations. Because of their nature, classical analyses based on mining techniques are consequently difficult to perform. Hence, the analysis of short texts poses several challenges for researchers.

Once a document collection is encoded via a bag-of-words scheme, every single term represents a different dimension in the vector space. The $terms \times documents$ table – obtained by juxtaposing the different document-vectors – is a large and very sparse matrix. If we deal with short texts, the problem of high dimensionality and sparsity is ever more severe. It is possible to cope with the data sparsity induced by the bag-of-words and amplified by the shortness of the texts in several ways. The simplest solution is to adapt existing techniques for standard documents, but there are also some proposal specifically designed for short texts. Some techniques take into account the relationship between the terms frequently occurring together in the texts (Seifzadeh et al., 2015; Yin & Wang, 2014). Starting from the $terms \times documents$ table, it is possible to derive a $terms \times terms$ table, in which the co-occurrence of linked terms allows to recover the information related to the context.

In order to reduce the dimensionality and enhance the knowledge extraction process, it is possible to consider high-level structures in the form of linked terms. This structures can be seen as concepts embodied in the document collection. As stated by psychologists, “concepts are the glue that holds our mental world together” (Murphy, 2004). Terms related to the same topic are likely to co-occur in the same text, thus they link together and form densely connected structures. On the $terms \times terms$ table, different approaches can be used for discovering and managing the emerging knowledge.

Within the heuristic optimisation approaches, *Ncut-weighted* non-negative matrix factorisation (Yan et al., 2012) tries to tackle the sparsity introducing a term weighting scheme based on co-occurrences. This approach measures term discriminability at a term level rather than considering traditional term weighting schemes which emphasise discriminability at a document level. Another proposal is the non-negative matrix factorisation on $term \times term$ correlation matrix (Cheng et al., 2013), that employs an alternative term correlation measure designed specifically for short texts. On the $term \times term$ correlation matrix, a symmetric non-negative matrix factorisation is performed to learn the topics and to assign them to the different documents subsequently. Similarly, in Seifzadeh et al. (2015), *generalised vector space model* is used to represent documents through an approximated $term \times term$ correlation matrix, obtained by selecting a few terms of the vocabulary with a quadrature method (Kumar et al., 2009).

A well-known alternative is the so-called *topic model*, based on a probabilistic standpoint of the problem. Topic modelling is a text mining technique used to uncover the underlying hidden topics or themes of a large collection of documents. Besides, it enables to group documents considering the thematic similarity. Starting from the *Latent Dirichlet Allocation* (LDA: Blei et al., 2003), many different techniques have been developed to cope with the issues of analysing short texts. *Biterm topic model* (Yan et al., 2013) directly models the term co-occurrence patterns to enhance the topic learning in the document collection. The Dirichlet multinomial mixture model-based approach proposed by Yin & Wang (2014) assigns a single topic to each document rather than considering a topic distribution. Lin et al. (2014) proposed a sparsity-enhanced topic model – the *dual-sparse topic model* – which modifies LDA to learn focused topics for each short text by replacing symmetric Dirichlet priors with Spike and Slab distributions (Ishwaran & Rao, 2005).

A common way of visualising the relationships between the terms belonging to a text is representing them as a graph (Carley, 1997; Popping, 2000). The main assumption is that both language and knowledge can be effectively modelled as a network of terms (Sowa, 1984). Several methodological proposals (e.g. Carley, 1988; James, 1992; Popping, 2003) and empirical studies (e.g. Fronzetti Colladon & Vagaggini, 2017; Gloor et al., 2017b; Shiau et al., 2017, 2018) were produced in this framework. More recent works (e.g. Misuraca et al., 2018) addressed the problem of information extraction from textual networks to the identification of high-level structures, considering a combination of terms as concepts or topics occurring in the collection of documents. Sayyadi & Raschid (2013) proposed the *KeyGraph* approach to topic detection, representing the document collection as a keyword co-occurrence graph. This method performs an off-the-shelf community detection algorithm based on *betweenness centrality* to group co-occurring keywords. The keywords belonging to each group are used to represent the topics. Each community of terms is represented as a feature vector, then the likelihood of the topic to document association is determined by cosine similarity. The clustering-based algorithm *ClusTop* (Lim et al., 2017) applies the Louvain community detection algorithm (Blondel et al., 2008) to the term network. Once topics have been identified, the algorithm assigns each document to the topic that has the highest co-occurrence of terms in both the document and the topic, weighting each term in the topic by its co-occurrence to the other terms. ClusTop has the advantage of automatically determining the appropriate number of topics, and it is also able to capture the syntactic meaning by using bigrams, trigrams and other term collocations. *WordCom* algorithm formulated by Jia et al. (2018) uses a term co-occurrence network weighted by the positive pointwise mutual information, performing the *k-rank-D* k-means type algorithm (Li et al., 2015) to identify semantic term communities. In the framework of probabilistic approaches, (Zuo et al., 2016) proposed a word network topic model based on the term co-occurrences. This method considers the distribution of terms over topics rather than the distribution of topics over documents, employing the standard Gibbs sampling for LDA to discover latent term groups (Henderson & Eliassi-Rad, 2009).

One of the main differences between topic modelling and community detection concerns the assignment of each term to a concept (Yang et al., 2013). Detecting concepts with topic modelling can be seen as a soft clustering approach, in which each term can belong to different clusters at the same time. Soft clustering associates a probability distribution with the term's membership to concepts, which means the more a term belongs to different concepts the less it is discriminant. Community detection can be seen as a hard clustering approach since the concepts are not overlapping and there is a high membership strength of each term to a single concept. Hard clustering associates an independent binary variable for each term and concept pair and, thus, do not suffer from the assumptions made by soft membership models.

Another aspect to consider is that topic modelling requires to prior set the number of concepts to be detected. Even if some methods for automatically setting the parameters have been proposed in the literature (e.g. Griffiths & Steyvers, 2004), there is a lack of shared solutions to cope with this problem. On the other hand, the majority of community detection algorithms allows to automatically determine the partition, since a hierarchical structure is implicitly assumed.

Here we propose a 2-fold strategy based on community detection and hierarchical clustering, to detect the main concept listed in a collection of short texts and categorise them on the basis of the knowledge emerging from the community structure. This heuristic approach allows to automatically explore big textual datasets retrieved from social media in an unsupervised perspective, offering to the management of a company a tool easy to implement and intuitive to use.

3. Problem definition and proposed strategy

As stated above, texts encode information in a form difficult to analyse from a quantitative point of view. Texts are seen initially as unstructured data and need a pretreatment for becoming structured data that can be processed with statistical techniques. At the same time, pretreatment is necessary to reduce language variability and eliminate the possible sources of noise, improving the effectiveness of the performed analyses (Song et al., 2005; Uysal & Gunal, 2014). Once the texts have been retrieved and parsed, it is possible to list in the so-called vocabulary all the terms used in the document collection. The first step of preprocessing is removing *stopwords* from the vocabulary, i.e. all the most common terms used in the language and the domain related to the analysed phenomenon. After terms have been filtered, it is possible to perform a normalisation in order to proceed uniformly in the whole collection of documents. Terms are converted to the same case (upper or lower), numbers are converted to their term equivalents or removed, and so on. The last step considers the inflexion of the terms and operates at a morphological level to reduce texts' variability. Two solutions are usually considered, *stemming* and *lemmatisation*. The main difference is that stemming reduces inflected terms to their roots by removing the affixes, whereas lemmatisation uses terms' lemma for reducing each inflected term to its canonical form. Even if stemmers are easier to implement and faster, it is preferable to consider lemmatisers in order to preserve the syntactic role of each term and improve the readability (Kettunen et al., 2005; Toman et al., 2006; Valbé et al., 2007; Konkol & Konopík, 2014).

Each pretreated document is transformed into a vector where the different elements represent the importance of each term according to the selected weighting scheme (Balbi & Misuraca, 2005). By juxtaposing the different document-vectors, a $terms \times documents$ table is built. One of the shortcomings of the vector space model is that it ignores the context in which terms are used. It is possible to get back part of the structural and semantic information by constructing a $terms \times terms$ co-occurrence table. Generally, each element of this table is the number of times two terms co-occur in the document collection, but other weighting schemes can also be used (e.g. Cheng et al., 2013). This data structure can be represented as a graph (V,E), where V is a finite set of nodes (or vertices) and E is a finite set of edges (or lines). Edges indicate the relationships between the nodes. We can build a network in which each term acts as a node and the co-occurrence between linked terms is expressed as an edge, visualising both single terms and subsets of terms frequently co-occurring together.

Highlighting if in a network there are groups of nodes sharing common characteristics, and/or playing similar roles within the graph, has several practical implications. The occurrence of groups of nodes that are more densely connected internally, suggests that the network has certain natural divisions within it. This means that the network can show at a local level some properties that are quite different from the ones showed at a global level, hence focusing on the whole structure may miss many interesting features concerning the analysed phenomenon. Literature refers to these groups of nodes as communities, even if there is a lack of a universally accepted definition. Several real-world networks concerning various applicative domains showed this kind of structures (e.g. Moody & White, 2003; Dourisboure et al., 2009; Papadopoulos et al., 2012; Antonacci et al., 2017). Communities are usually thought as subgraphs densely inter-connected and sparsely connected to other parts of the network (Wasserman & Faust, 1994). From a theoretical viewpoint, communities are then not very different from clusters (Fortunato & Hric, 2016).

3.1. Basic notation and data structure

Let $\mathbf{T} = \{\mathbf{d_1}, \ldots, \mathbf{d_n}\} \subset \Re^w$ be a set of n document vectors in a term space of dimension w. This set can be represented as a $terms \times documents$ table, where each element t_{ij} represents the number of occurrences of a term i into a document j $(i = 1, \ldots, w; j = 1, \ldots, n)$.

Let transform **T** into a binary matrix **B**, where the generic element b_{ij} is equal to 1 if the term i occurred at least once in document j, 0 otherwise. From the matrix **B**, we derive the $terms \times terms$ co-occurrence matrix **A** by the matrix multiplication $\mathbf{BB}^\mathsf{T}$. The generic element $a_{ii'}$ is the number of documents in which the term i and the term i' co-occur $(i \neq i')$. According to network theory, **A** is a $w \times w$ undirected weighted adjacency matrix that can be used to visualise the relations existing among the different nodes. The a_{ii} elements on the main diagonal of **A** – known as loops – represents the number of documents containing each term i. In the follows they are set to 0 since they are not informative for the purpose of our strategy. Figure 1 shows the data structure described above.

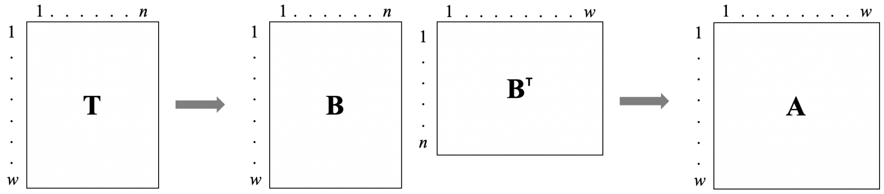

Figure 1: Data structure: from the $terms \times documents$ table **T** to the $terms \times terms$ co-occurrence table **A**

3.2. Network-based concept extraction

It is possible to perform on the matrix **A** a community detection procedure, with the aim of extracting from the document collection the mostly occurring concepts. Different algorithms have been proposed for this task. Traditional approaches are based on the well-known hierarchical or partitional clustering (Scott, 2000). The main difference is that in hierarchical clustering inter-cluster edges are removed, whereas in partitional clustering edges between pairs of nodes showing low similarity are removed. Divisive approaches do not introduce substantial conceptual advances with respect to traditional ones.

The most popular algorithm for community detection was developed by Newman & Girvan (2004). This proposal introduced the concept of *modularity* as a stopping criterion. Modularity is the difference between the observed fraction of edges that fall within the given communities and the expected fraction in the hypothesis of random distribution. Let suppose that the nodes of matrix **A** can be divided into two communities. The membership to one community or the other one is detected by a variable s, assuming values 1 or -1 respectively. The modularity Q is defined as:

$$Q = \frac{1}{2h}\sum_{ii'}\left[a_{ii'} - \frac{\delta_i \delta_{i'}}{2h}\right] s_i s_{i'} \tag{1}$$

where δ_i is the degree of the i-th term, h is the total number of edges in the network, and s_i represents the membership value of the term i to a community. When we consider G communities, eq. 1 can be expressed in terms of additional contribution ΔQ to the modularity. In matrix form we have:

$$\Delta Q = \frac{1}{4h}\mathbf{s}^{\top}\mathbf{M}^{(g)}\mathbf{s} \tag{2}$$

where $\mathbf{M}^{(g)}$ is the modularity matrix referred to the g-th community $(g = 1, \ldots, G)$, and $\mathbf{s}$ is the binary column vector indicating the membership of each term to the community. The value of modularity Q ranges in an interval $[0;1]$, where 0 indicates a random structure and 1 indicates strong community structure (Newman, 2003). From an empirical viewpoint, it has been observed that modularity values usually fall in a subinterval $[0.3;0.7]$ (Newman & Girvan, 2004).

Starting from the method of Newman & Girvan, several algorithms based on modularity have been proposed for community detection (e.g. Newman, 2006; Pons & Latapy, 2006; Rosvall et al., 2009). In our strategy, a greedy approach is used for community detection (Clauset et al., 2004). The *fast-greedy* algorithm falls in the general family of agglomerative hierarchical clustering methods. The advantage of using this approach is that the problem of choosing a grouping criterion is overcome by the direct use of modularity as optimisation function. This algorithm starts with a state in which each term is the sole member of one of the G communities, then repeatedly joins communities together in pairs choosing in each step the join that results in the greatest increase in modularity.

At the end of the detection process, it is possible to build a $terms \times concepts$ table $\mathbf{C}$. This table is a complete disjunctive matrix where the c_{ig} element is 1 or 0 when a term i belongs or not belongs to a concept. The information contained in $\mathbf{C}$ can be then used for obtaining a $documents \times concepts$ table $\mathbf{T}^{\star} \equiv (\mathbf{T}^{\top}\mathbf{C})\mathbf{D}_G{}^{-1}$, where $\mathbf{D}_G{}^{-1}$ is the diagonal matrix obtained from the column marginal distribution of $\mathbf{C}$. Each cell of $\mathbf{T}^{\star}$ contains the proportion of terms belonging to a concept. Aiming at grouping the documents based on the concepts, any clustering algorithm can be performed on $\mathbf{T}^{\star}$. In particular, we propose to use Ward's agglomerative clustering algorithm. This method is the only one among agglomerative clustering that is based on a sum-of-squares criterion, producing groups that minimise within-cluster variability at each aggregation (Murtagh & Legendre, 2014). The advantage of using this approach relies on the possibility of performing clustering in an unsupervised framework. Having an explorative perspective, it is then not necessary to consider any prior knowledge on the document collection. Moreover, the clustering procedure is designed to work parameter free. By obtaining a sequence of nested partitions, it is not necessary to set-up the algorithm with a given number of clusters or to repeat the procedure by considering different alternative solutions.

Once the documents are grouped, they can be routed to the different support areas for helping customers and give a quick reply on the social media platform. Figure 2 graphically shows how a system based on the proposed strategy would work.

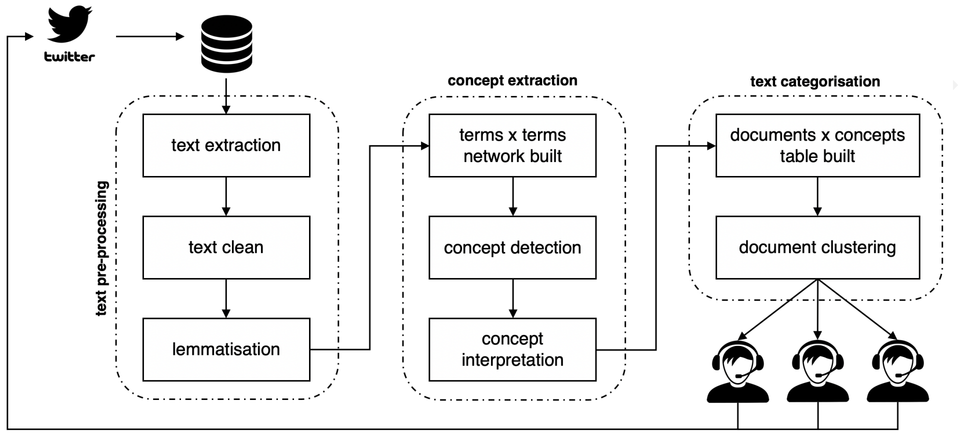

Figure 2: Architecture of a system based on the proposed strategy

In the following, the effectiveness of our strategy is showed by analysing a set of tweets posted by customers in the AppleSupport official account on Twitter. The different steps have been implemented by using R (R Core Team, 2013).

4. A case study: the AppleSupport service on Twitter

Twitter is one of the most popular – and worldwide leading – social networking service. It can be seen as a blend of instant messaging, microblogging and texting, with brief content and a vast audience. The embryonic idea was developed considering the exchange of texts like Short Message Service in a small group of users. Users share opinions on a variety of topics, discuss current issues and write about their experiences. Each tweet may contain texts, hashtags, usernames, links, emojis, pictures, videos, and is limited to 280 characters. As of the third quarter of 2017, it has 330 million monthly active users, with an amount of daily sent tweets close to 500 million (Source: *Twitter*, *Statista*).

Twitter has become a valuable source of data that can be efficiently used for different domains and applications. Such data are mainly used for marketing and social studies (e.g. Shirdastian et al., 2017; Aswani et al., 2018), but research interests on Twitter data cover the most diverse fields and range from identifying the drivers of voter behaviour (Grover et al., 2018) to emergency management activities (Singh et al., 2017). Since 2016, Twitter has introduced new functionalities designed explicitly for customer care. *Direct Messages* are a way for customers to have private conversations with a company. Companies can add a deep link to their tweets automatically displaying a call to action button, which allows the customer to send the business a direct message. Moreover, Twitter introduced *Customer Feedback*, enabling people to privately share their opinions with a business after a service interaction. For this reason, the micro-blogging platform became the primary social media for customer service.

4.1. Data description and pre-treatment

In order to explore the effectiveness of our proposal, we downloaded from Kaggle a dataset related to customer support on Twitter[1]. The Kaggle's *Customer Support on Twitter* dataset is a large, modern corpus of tweets useful for studying the customer support practices and impact. It contains over 3 million tweets and replies from the biggest brands – i.e. AmazonHelp, SpotifyCares, AppleSupport – and offers conversations between customers and customer support agents on Twitter. It has different important advantages over other conversational text datasets. Customers contact support services to solve a specific problem, and the manifold of problems to be discussed is relatively small, especially compared to other unconstrained conversational datasets like the *reddit* corpus used by other authors (Jeong et al., 2017). Moreover, the brevity of tweets causes more natural responses from support agents and to-the-point descriptions of problems and solutions.

We considered only the tweets sent to the customer support of Apple (*@AppleSupport*). Apple launched a Twitter account dedicated to customer support only in March 2016. For years, Apple has been held up as one of the top companies for customer service, having its own format for customer service in the stores (the *Genius Bars*), but decided to use social media late with respect to other competitors. The tweets are published between September 27th and December 1st 2017. We extracted only the messages written in English. The pre-processing was performed in two steps. First, we stripped URLs, usernames, hashtags, emoticons, and we normalised the tweets by removing special characters and any separators than blanks. Second, on the cleaned tweets, we performed a lemmatisation and

[1] https://www.kaggle.com/donyoe/exploring-customer-support-in-twitter/data

a grammatical tagging. In the analysis, we consider only nouns and adjectives because of their content-bearing role. Moreover, we delete from the vocabulary the terms occurring less than 2 times and also the terms long less than 3 characters. Thus we obtain a *terms* × *documents* table **T** with 2214 rows and 106860 columns, and a *terms* × *terms* co-occurrence table **A**. In Figure 3 the 100 most occurring terms belonging to the collection are displayed as a word cloud. It is possible to see that the customers mostly used terms related to the Apple mobile devices, especially considering the operating system and other software problems.

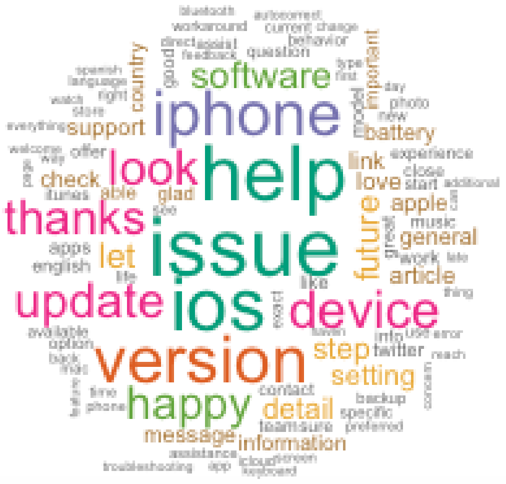

Figure 3: Word cloud of the 100 most used terms in AppleSupport dataset

4.2. Concept identification and categorisation process

We performed the community detection procedure on **A** to identify the different concepts listed in the collection. In order to prune the network from isolated and peripheral terms, a threshold $\hat{a}$ on the co-occurrences can be set. As a consequence, from a computational viewpoint, removing infrequent collocations speeds up the detection process. Table 1 shows how different threshold values affect the network structure. Following (Fronzetti Colladon & Gloor, 2018), we calculated several metrics to evaluate the stability of the whole network: the *average distance among reachable pairs* (ADARP), the *average degree*, the *clustering coefficient* (CC), the *density* and the *diameter*.

As we can see, filtering and deleting isolated and peripheral terms do not have a big impact on the whole network structure. At different threshold, the diameter remained unchanged and the other metrics showed small variations. Looking at the percentage of original information saved for different threshold values, with $\hat{a} = 5$ we retained the 50.22% of

Table 1: Network metrics for different co-occurence thresholds

â	% of nodes	ADARP	Avg. degree	CC	Density	Diameter
1	91.15	2.302	55.637	0.302	0.028	5
2	73.35	2.333	45.764	0.307	0.028	5
3	61.97	2.324	41.335	0.309	0.030	5
4	55.28	2.335	37.670	0.309	0.031	5
5	50.22	2.343	35.094	0.310	0.032	5
6	46.12	2.339	33.030	0.308	0.032	5
7	42.46	2.340	31.736	0.310	0.034	5
8	40.11	2.357	30.243	0.311	0.034	6
9	38.44	2.381	28.761	0.310	0.034	6
10	36.31	2.374	27.930	0.311	0.035	5
11	34.60	2.369	27.081	0.311	0.035	5
12	33.24	2.378	26.223	0.311	0.036	6

the terms listed in **A**. By applying the greedy algorithm on the sub-network obtained after the pruning, we detected 26 different concepts. The high value of the modularity measure ($Q = 0.713$) supports the effectiveness of our strategy.

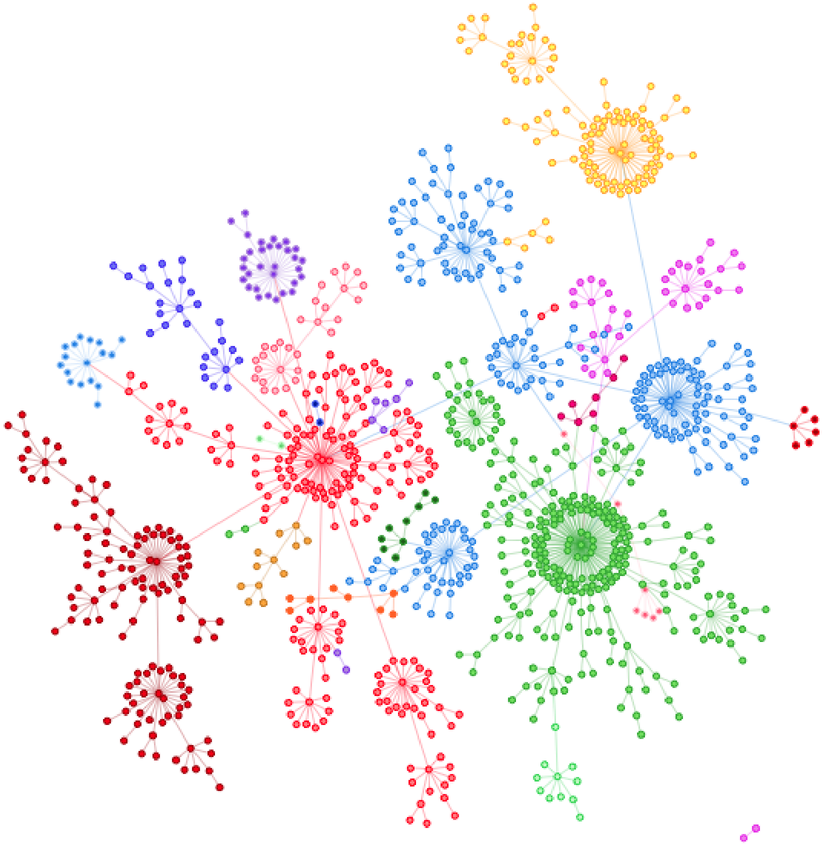

Figure 4: Communities/concepts detected on the term network of AppleSupport tweets

Figure 4 shows in different colours the communities/concepts detected in the network (with some colours being reused due to the large number of communities yielded by the procedure). It is possible to see that some concepts are described by few terms (e.g. the community depicted in pink on the bottom-right side of the network), whereas other concepts are described by a huge number of terms (e.g. the community depicted in green on the right side of the network and the community depicted in garnet-red on the left side of the network). In order to better explore the concepts, Table 2 shows an example of the terms belonging to the first seven concepts.

Table 2: First seven concepts and corresponding terms detected by the proposed strategy

Concept	**Nodes**	**Terms**
C1	201	*version, ios, setting, backup, alarm, siri, keyboard, reset, . . .*
C2	193	*help, detail, thanks, good, like, assistance, information, . . .*
C3	104	*internet, safari, podcast, firmware, wifi, modem, connection, . . .*
C4	268	*model, home, button, battery, life, screen, low, power, charge, . . .*
C5	27	*offer, support, team, spanish, english, contact, . . .*
C6	119	*app, store, music, song, apple, account, downloaded, play, . . .*
C7	31	*message, error, imessage, sms, direct, send, post, . . .*

It is possible to note that concept **C1** refers to the operating system IOS and other software problems, **C4** refers to the iPhone, its battery life and other hardware problems, **C6** refers to the Apple Music service, and so on. The results show that with the community detection procedure it is possible to discriminate different requests from the customers concerning the Apple products and services.

4.3. A comparison with topic modelling

To show the effectiveness of the proposed strategy, we performed on the Kaggle dataset a topic modelling method specifically developed for short texts, applying the Biterm topic model (BTM) proposed by Yan et al. (2013). This approach considers the co-occurrence of the terms used in the collection by looking at bigrams. It means that the terms are linked in a local context only if they are adjacents, without taking into account the whole text.

As stated above, there is not a shared procedure to automatically set the number of topics in the model. We decided to consider as a reference the partition in 26 concepts obtained by the community detection step of our strategy. Since BTM produces a term-topic probability distribution, every single term of the vocabulary has a probability of belonging to the different topics. We filtered the term distribution for each topic by setting a threshold of 0.01 and saved only the top terms characterising the topics themselves. It is important to point out that topic modelling produces soft solutions; hence each term can appear in more than one topic. Table 3 shows the first seven topics detected by BTM and the correspondent communities detected by the proposed strategy.

Table 3: First seven topics and corresponding terms detected by BTM

Topic	N.of terms	Terms	Concept
S1	17	*ios, update, version, device, issue, late, assist, release, backup,....*	-
S2	20	*ios, version, help, issue, device, experience, look, detail,....*	-
S3	24	*country, locate, message, option, look, information, start,.....*	C10
S4	19	*battery, life, important, help, charge, iphone, device, ios,...*	C4
S5	17	*device, ios, help, use, version, issue, mac, apple, work,.....*	-
S6	23	*help, issue, support, experience, detail, team, link, reach,.....*	-
S7	21	*update, ios, help, general, issue, step, device, backup,.....*	-

Looking at the topic compositions, we noted that the most common terms (e.g. *ios*, *version*, *help*, *device*) had a high probability to belong to different topics. On the other hand, they seemed to be not discriminant in describing specific concepts related to the requests and the complaints of the customers contacting the AppleSupport service. Table 4 shows in more detail some communities having a corresponding topic detected by BTM. It is possible to see that the concepts obtained by community detection are more informative and coherent than the topics.

4.4. Text categorisation

Once we obtained the concepts, we categorised the texts on the basis of this new knowledge. By selecting only the terms belonging to the different communities, we obtained a 103659×26 table $\mathbf{T}^{\star}$. There were 3201 tweets not containing any of the detected concepts since we filtered terms with low co-occurrence.

On this new table, we performed a hierarchical clustering based on the Ward criterion. Figure 5 shows the histogram of the within-cluster sum of squares calculated for each partition of the clustering procedure, representing in this way the loss of intra-class variability caused by the aggregation. The maximum gap in the distribution suggests considering a partition in 25 clusters.

Because of the unsupervised nature of the approach, the quality of the results can be investigated only by looking at the cluster composition. Due to the limitation of 140 characters, each tweet can express one to three concepts at most. Table 5 shows the concepts

Table 4: Comparison between concepts and topics

Concept	**Topic**
C4: *model, home, button, battery, life, screen, low, power, charge, anxious, charger, liquid, adapter, switch, light, recycling, usb, temperature, warning, hot, warm, slow*	**S4**: *battery, life, important, help, charge, iphone, device, ios, work, let, version, look, use, issue, start, last, reset, step*
C5: *offer, support, twitter, english, spanish, available, team, contact, preferred, language, guidance, advisor, reach, much, imovie, sale, developer, department, logic, inquiry, sales, phony, xcode, airport*	**S21**: *support, twitter, english, help, language, spanish, available*
C7: *message, group, error, imessage, sms, mms, direct, send, hasn, post, someone, character, box, green, imessages, private, receiving, individual, thread, junk, spam, reading, people, arrow, blue, gif, recipient, forwarding, troublesome*	**S11**: *message, help, error, email, use, attempt, type, device, able, send, issue, look, ios, report, link, emails, article*
C10: *country, option, service, globe, located, region, provider, repair, residence, authorized, continuity, damaged*	**S3**: *country, locate, message, option, look, information, start, direct, detail, help, gather, issue, good, region, troubleshooting, use, link, apple, iphone, service,device, let*

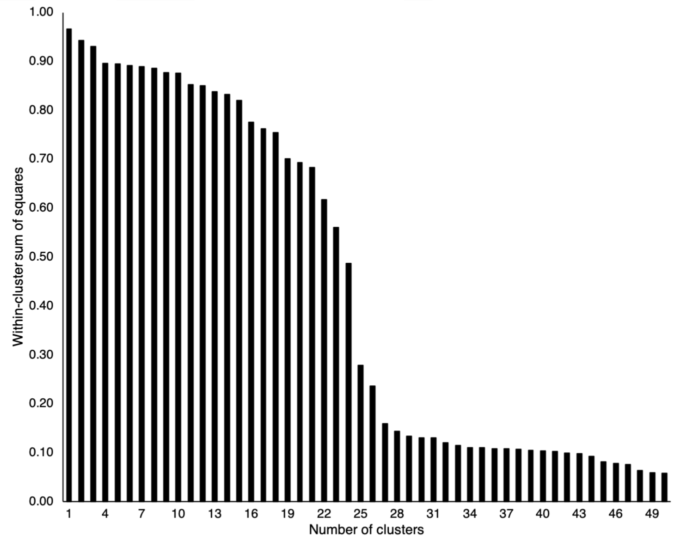

Figure 5: Histogram of the within-cluster sum of squares calculated on the different partitions

characterising the different clusters. The order of the concepts represents their importance in terms of statistical significance. According to Nakacha & Confais (2004), it is possible to consider a variable X_{kg} computing the number of tweets in a given cluster k that contain a concept g. This variable follows a hypergeometric distribution and can be used as test statistics for evaluating the characterisation of the clusters. The more we observe a high value x_{kg} of the test statistics, the more the corresponding p-value $p << 0.001$ indicating that the concept is peculiar of that given cluster.

Table 5: Clusters'size and composition

Cluster	**N. of tweets**	**Concepts (test statistics[†])**
1	21592	**C1** (182.32),**C4** (138.90)
2	41483	**C2** (122.68)
3	5995	**C6** (244.55)
4	4020	**C8** (258.11)
5	2625	**C9** (272.65)
6	4163	**C10** (274.59)
7	3657	**C7** (275.58)
8	621	**C20** (269.48)
9	1295	**C19** (282.33)
10	1260	**C11** (284.42)
11	1091	**C24** (296.20)
12	8820	**C3** (273.88)
13	877	**C15** (298.78)
14	428	**C14** (293.02)
15	649	**C17** (296.51)
16	1081	**C16** (300.28)
17	2492	**C5** (293.39)
18	68	**C23** (303.67)
19	496	**C12** (297.19)
20	505	**C18** (300.63)
21	84	**C22** (303.96)
22	33	**C21** (305.05)
23	177	**C26** (308.69)
24	50	**C25** (311.86)
25	7	**C13** (316.28)
TOT	103569	

[†] *Test statistics* $n_{kg} > 50$*, p-value* $p << 0.001$

5. Discussion

In this paper, we introduce a two-fold strategy aiming at automatically extracting and organising the information contained in the written requests to social customer care services of companies. The approach is based on a community detection step to highlight the main issues as high-level structures of terms expressing different concepts. Successively, the texts are clustered via an unsupervised method by using this new knowledge base. We showed the effectiveness of the proposal on a document collection encompassing the 106,860 requests sent in October-November 2017 to the AppleSupport account on Twitter.

Several theoretical and practical implications can be derived from our research findings.

5.1. Theoretical implications

Text mining aims at analysing textual data in an attempt to discover patterns and implicit meanings hidden within texts. Since documents can be seen as unstructured data, a pre-treatment is necessary to perform quantitative processing. One of the main problems in transforming a document collection into a matrix concerns the high dimensionality and sparsity, once texts have been encoded via the bag-of-words scheme and represented as vectors. The two aspects are related to the vocabulary of terms used in the collection, including both content-bearing terms as well as stopwords and common terms without any discriminative power. As pointed out by several authors, the analysis and the interpretation of such datasets is critical (e.g. Steinbach et al., 2003). Moreover, reading and understanding the results is not easy because there are no references to the contexts in which terms have been used. All these disadvantages are enhanced when short texts are considered. The analysis of short texts retrieved from social media platforms has proved to be a hot topic in the most recent developments of mining techniques. The strategy we propose relies on a different organisation of data, starting from a $terms \times terms$ co-occurrence table that can be easily displayed as a network. This kind of representation retains the context information of texts, with higher readability of the contents with respect to other solutions (Jin & Srihari, 2007; Zhou et al., 2010). At the same time, network representation allows reducing sparseness by filtering and deleting the isolated terms.

Aiming at extracting knowledge with respect to a given domain, we considered the advantage of highlighting high-level textual structures in the form of concepts, instead of using the informative power of single terms. Network tools allow operating on the $terms \times terms$ co-occurrence table, considering a community detection procedure to discover concepts. This approach enables to define communities of linked terms similarly to hard clustering, by assigning each term only to one community on the basis of the co-occurrences among different terms. Mikhina & Trifalenkov (2018) showed the effectiveness of an analogous framework for text clustering so that working on the terms side can produce good results too. We showed that other approaches like BTM produces less accurate results. Assigning terms to more communities as in soft clustering approaches does not seem to define the different concepts better. Moreover, community detection does not depend on any input parameters, where probabilistic methods require careful tuning.

The second step of our strategy concerns the categorisation of the document belonging to the collection on the basis of the new knowledge base represented by the concepts. Since the community detection is applied on the terms, hiding the relations among the documents, it is necessary to cross-tabulate these latter with the concepts. We introduced a new $documents \times concepts$ table, expressing the intensity of each concept in each document as the proportion of terms belonging to a concept that is used in the document itself. The idea of categorising documents on the basis of concepts is not new, and several authors demonstrated the effectiveness of this approach with respect to the more traditional frameworks based on the terms (e.g. Chen et al., 2010; Shehata et al., 2010; Fu et al., 2011). In our proposal, an unsupervised approach based on hierarchical clustering is used, since we designed the overall process to be data-driven and automatic. It is necessary to underline that our strategy is a heuristic approach to the problem, offering a quick solution that is easy to understand and implement. Heuristics are practical, serving as fast and feasible short-term solutions to planning and scheduling problems. On the contrary, a highly sophisticated process requires specific expertise and technologies.

Even if the proposal is here tailored for the specific problem of social customer care, we experienced the effectiveness of this two-fold approach on different contexts (Balbi et al., 2018; Misuraca et al., 2018). Analogously, it is possible to apply the strategy to short texts from other social media like Facebook or Reddit. The only difference relies on the scraping step, according to the social media policies and the functionality of the corresponding APIs. Concerning the use of the strategy for long-term texts, it is necessary to consider how co-occurrences are calculated. A document-level term co-occurrence leads to a coarse-grained representation than a sentence-level term co-occurrence. Some limitation and future development of this research are discussed in the last section of this paper.

5.2. Implications for managers and practitioners

In recent years, marketers are expanding companies' strategies to design and manage the entire process customers go through so that having a good experience become productive for the companies themselves. Collecting and analysing customer feedbacks is then an essential key to success since it allows companies to learn continuously and adapt their offerings to customer preferences and needs (Sun & Shibo, 2011). This task is more and more strategic since companies that analyse customer data regularly are averagely 5% more productive and 6% more profitable than their competitors (Vidgen et al., 2017). The literature recognised that customer evaluation of service experiences are an outcome of the interactions among companies, employees and customers (Voorhees et al., 2017), indicating that customer service can be seen as a *value co-creation process* (Vega-Vazquez et al., 2013). Since customers prefer to have direct links with companies, many organisations are improving their presence on social media (Chung et al., 2017). Posting a comment on companies' official social media accounts allows customers to ask help more freely, looking at a rapid solution, and sharing the most critical situations they experienced with the overall community of customers and potential customers. It has been shown that adopting social media enhance business values, increasing customer loyalty and retention, increasing sales and revenues, improving customer satisfaction, creating brand awareness and building reputation (e.g. Laroche et al., 2013; Fronzetti Colladon, 2018).

Unstructured data are increasingly considered alongside classically structured data, making harder for companies to implement efficient and effective processes to manage all the information. Text mining methods offer a potential solution for dealing with the huge volumes of unstructured data in a business domain (Ur-Rahman & Harding, 2012; Singh et al., 2016). The deployment of text mining models has clear managerial implications, including the availability of accurate and timely information, for better-informed decision making. In this framework, companies used both manual or automatic approaches. Many companies formed a dedicated customer service team, consisting of several human agents trained to support customers in their various needs. Social customer care requires to carefully establish the strategy to be used, in terms of skills to be distributed within the service, promptness of response, netiquette, and so on (Gloor et al., 2017a). Strategies based on a manual approach can lead to a deeper understanding of customer requests but tend to be inconsistent when vast quantities of data have to be reviewed. Moreover, this manually addressing requests is time-consuming and often fails customers' expectations. On the other hand, running automated strategies can fail to achieve companies' expectations because of the poor flexibility in adapting to different situations.

In an operational perspective, our strategy offers a fast and simple solution for monitoring the needs of customers, addressing their requests to the members of the customer care team that can offer in real time the most suitable solution. Firstly, the application of the strategy proposed in this paper enables both to reduce the complexity of textual datasets and retain higher readability of the results with respect to the analysed context. One of the main advantages is that the procedure is data-driven, hence it is not necessary to use prior or additional information. Moreover, the procedure is designed to be unsupervised, so that an operator does not have to choose any parameter to carry on the analysis. The only intervention required concerns the pre-treatment of the text, but it is possible to automatise also this step. Secondly, the strategy can be easily implemented in a business intelligence system and used on different levels in the company (Subramaniam et al., 2009). Several studies showed the benefits of automating the analysis of large amounts of customer textual data related to different products and services, demonstrating how to effectively manage and convert customers' feedbacks to help business strategies (Khare & Chougule, 2012; Ur-Rahman & Harding, 2012; Ludwig et al., 2013). Marketers and managers can use the strategy to highlight the strengths and weakness of a product, by looking at its usage by customers (Chung & Tseng, 2012). Segmentation of customers sending requests via social media can also be derived from meta-data (e.g. Sloan et al., 2015; An et al., 2018), to deeply understand which kind of customers asks assistance and for which kind of problems. Technical support can prepare a knowledge base, with the most suitable solutions to the different problems emerging from the analysis. Moreover, by routing the requests of each customer to the operators of the social customer care, on the basis of the categorisation of the different requests, more effective assistance can be offered to the customers. More prominent organisations can also implement chat-bots instead of human operators in a more complex architecture (Xu et al., 2017), limiting costs and reducing the time customers have to wait for a solution.

6. Conclusion and future research

The use of social media and new communication tools changed the way customers interact with companies. This (r)evolution pushed the organisations to rethink the customer care services. Our strategy offers an automatic framework to support customers' needs and complaints sent via social media, easy to implement in a business intelligence system. Even if the preliminary results seem to be very promising, a deep investigation has to be considered to validate the proposal. In particular, the use of benchmark collections has to be considered for an accurate quantitative evaluation of our strategy.

Some limitations have to be pointed out, but they provide interesting directions for further research. Firstly, we set the co-occurrence threshold in the community detection procedure by looking at the network structure. The value we chose in the analysis ($\hat{a} = 5$) removed terms infrequently co-occurring in the document collection. This practical consideration greatly sped up the community detection. However, we noted that this parameter was largely unoptimised and future work may benefit from a different accounting for paired terms' weights, e.g. by considering other measures of term-term association. Secondly, we implicitly considered the collection and the corresponding network as static. In a real situation, new requests arrive continuously. Having this perspective, a dynamic framework should be designed. Several clustering algorithms can be applied, but particularly interesting could be the use of supervised techniques based on the information coming from prior analyses on the same domain, in a lifelong learning perspective (Thrun, 1998; Silver et al., 2013). On the best of our knowledge, this is the first attempt of using text mining techniques and network analysis in a social media care framework. This research domain offers several open challenges (Stieglitz et al., 2018), suggesting more and more future contributions and developments both from a statistical and computational viewpoint.

References

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

An, J., Kwak, H., Jung, S.-g., Salminen, J., & Jansen, B. J. (2018). Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, *8*. URL: `https://doi.org/10.1007/s13278-018-0531-0`.

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*, 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*, *81*, 107–116.

Balbi, S., & Misuraca, M. (2005). Visualization techniques in non-symmetrical relationships. In S. Sirmakessis (Ed.), *Knowledge mining* (pp. 23–29). Springer–Verlag.

Balbi, S., Misuraca, M., & Spano, M. (2018). A two-step strategy for improving categorisation of short texts. In D. F. Iezzi, L. Celardo, & M. Misuraca (Eds.), *Proceedings of 14th International Conference on Statistical Analysis of Textual Data (JADT18)* (pp. 60–67). Universitalia volume 1.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*. URL: `http://stacks.iop.org/1742-5468/2008/i=10/a=P10008`.

Carley, K. (1988). Formalizing the social expert's knowledge. *Sociological Methods and Research*, *17*, 165–232.

Carley, K. (1997). Network text analysis: the network position of concepts. In C. Roberts (Ed.), *Text analysis for the social sciences* (pp. 79–102). Lawrence Erlbaum Associates.

Chen, L., Jie, Z., & Bi-Cheng, L. (2010). A text categorization framework based on concept structure. In *2nd International Conference on Computer Engineering and Technology* (pp. 569–573). volume 3.

Cheng, X., Guo, J., Liu, S., Wang, Y., & Yan, X. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the 13th SIAM International Conference on Data Mining* (pp. 749–757).

Chung, A., Andreev, P., Benyucef, M., Duane, A., & O'Reilly, P. (2017). Managing an organisation's social media presence: An empirical stages of growth model. *International Journal of Information Management*, *37*, 1405–1417.

Chung, W., & Tseng, T.-L. (2012). Discovering business intelligence from online product reviews: A rule-induction framework. *Expert Systems with Applications*, *39*, 11870–11879.

Clauset, A., Newman, M. E. J., & Cristopher, M. (2004). Finding community structure in very large networks. *Physical Review E*, *70*.

Dourisboure, Y., Geraci, F., & Pellegrini, M. (2009). Extraction and classification of dense implicit communities in the web graph. *ACM Transactions on the Web*, *3*. URL: `http://doi.org/10.1145/1513876.1513879`.

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*.

Fronzetti Colladon, A. (2018). The semantic brand score. *Journal of Business Research*, *88*, 150–160.

Fronzetti Colladon, A., & Gloor, P. (2018). Measuring the impact of spammers on e-mail and twitter networks. *International Journal of Information Management*, *(in press)*.

Fronzetti Colladon, A., & Vagaggini, F. (2017). Robustness and stability of enterprise intranet social networks: the impact of moderators. *Information Processing and Management*, *53*, 1287–1298.

Fu, X., Liu, L., Gong, T., & Tao, L. (2011). Improving text classification with concept index terms and expansion terms. In D. Liu, H. Zhang, M. Polycarpou, C. Alippi, & H. He (Eds.), *Advances in Neural Networks – ISNN 2011* (pp. 485–492). Springer–Verlag.

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

Gloor, P., Fronzetti Colladon, A., Grippa, F., & Giacomelli, G. (2017b). Forecasting managerial turnover through e-mail based social network analysis. *Computers in Human Behavior*, *71*, 343–352.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *101*, 5228–5235.

Grönroos, C. (1978). A service-oriented approach to marketing of services. *European Journal of Marketing*, *12*, 588–601.

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2018). Polarization and acculturation in the 2016 US presidential election: Can twitter analytics predict changes in voting preferences? *Technological Forecasting and Social Change*, . URL: `https://doi.org/10.1016/j.techfore.2018.09.009`.

Heller Baird, C., & Parasnis, G. (2011). From social media to social crm: reinventing the customer relationship. *Strategy & Leadership*, *39*, 27–34.

Henderson, K., & Eliassi-Rad, T. (2009). Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM Symposium on Applied Computing* (pp. 1456–1461).

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, *33*, 730–773.

James, P. (1992). Knowledge graphs. In R. van der Riet, & R. Meersman (Eds.), *Linguistic Instruments in Knowledge Engineering* (pp. 97–117). Elsevier.

Jeong, B., Yoon, J., & Lee, J. (2017). Social media mining for product planning: a product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, *(in press)*.

Jia, C., Carson, M. B., Wang, X., & Yu, J. (2018). Concept decompositions for short text clustering by identifying word communities. *Pattern Recognition*, *76*, 691–703.

Jiang, L., Jun, M., & Yang, Z. (2016). Customer-perceived value and loyalty: how do key service quality dimensions matter in the context of b2c e-commerce? *Service Business*, *10*, 301–317.

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.

Jin, W., & Srihari, R. K. (2007). Graph-based text representation and knowledge discovery. In *Proceedings of the 2007 ACM symposium on Applied computing* (pp. 807–811).

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, *53*, 59–68.

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

Karakaya, F., & Ganim Barnes, N. (2010). Impact of online reviews of customer care experience on brand or company selection. *Journal of Consumer Marketing*, *27*, 447–457.

Kettunen, K., Kunttu, T., & Järvelin, K. (2005). To stem or lemmatize a highly inflectional language in a probabilistic IR environment. *Journal of Documentation*, *61*, 476–496.

Khare, V. R., & Chougule, R. (2012). Decision support for improved service effectiveness using domain aware text mining. *Knowledge-Based Systems*, *33*, 29–40.

Konkol, M., & Konopík, M. (2014). Named entity recognition for highly inflectional languages: Effects of various lemmatization and stemming approaches. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Proceedings of the 17th International Conference on Text, Speech and Dialogue* Lecture Notes in Computer Science (pp. 267–274).

Kumar, S., Mohri, M., & Talwalkar, A. (2009). Sampling techniques for the nyström method. In D. van Dyk, & M. Welling (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (pp. 304–311).

Laroche, M., Habibi, M. R., & Richard, M.-O. (2013). To be or not to be in social media: How brand loyalty is affected by social media? *International Journal of Information Management*, *33*, 76–82.

Li, Y., Jia, C., & Yu, J. (2015). A parameter-free community detection method based on centrality and dispersion of nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, *438*, 321–334.

Lim, K. H., Karunasekera, S., & Harwood, A. (2017). Clustop: A clustering-based topic modelling algorithm for twitter using word networks. In J.-Y. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. A. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, & M. Toyoda (Eds.), *Proceedings of the 2017 IEEE International Conference on Big Data* (pp. 2009–2018).

Lin, T., Tian, W., Mei, Q., & Cheng, H. (2014). The dual-sparse topic model: Mining focused topics and focused terms in short text. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 539–550).

Ludwig, S., de Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, *77*, 87–52.

Mikhina, E. K., & Trifalenkov, V. I. (2018). Text clustering as graph community detection. *Procedia Computer Science*, *123*, 271–277.

Misuraca, M., Scepi, G., & Spano, M. (2018). A network approach to dimensionality reduction in text mining. In A. Abbruzzo, E. Brentari, M. Chiodi, & D. Piacentino (Eds.), *Book of Short Papers SIS 2018* (pp. 344–351). Pearson.

Moody, J., & White, D. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, *68*, 103–127.

Murphy, G. (2004). *The Big Book of Concepts*. MIT press.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, *31*, 274–295.

Nakacha, J., & Confais, J. (2004). *Approche pragmatique de la classification*. Editions TECHNIP.

Newell, F. (2001). *Loyalty.com: Customer Relationship Management in the New Era of Internet Marketing*. McGraw-Hill.

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, *67*.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, *74*, 1–19.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*.

Papadopoulos, S., Kompatsiaris, I., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knwoledge Discovery*, *24*, 515–554.

Pinto, D., Rosso, P., & Jiménez-Salazar, H. (2010). A self-enriching methodology for clustering narrow domain short texts. *The Computer Journal*, *54*, 1148–1165.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of graph algotithms and apprlications*, *10*, 191–218.

Popping, R. (2000). *Computer-assisted Text Analysis*. Sage Publications.

Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information*, *42*, 91–106.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna. URL: `http://www.r-project.org/`.

Rosenbaum, M., & Massiah, C. (2007). When customers receive support from other customers. exploring the influence of intercustomer social support on customer voluntary performance. *Journal of Service Research*, *9*, 257–270.

Rosvall, M., Axelsson, D., & Bergstrom, C. (2009). The map equation. *The European Physical Journal Special Topics*, *178*, 13–23.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*, 613–620.

Sandes, F. S., & Urdan, A. T. (2013). Electronic word-of-mouth impacts on consumer behavior: Exploratory and experimental studies. *Journal of International Consumer Marketing*, *25*, 181–197.

Sayyadi, H., & Raschid, L. (2013). A graph analytical approach for topic detection. *ACM Transactions on Internet Technology*, *13*, 1–23.

Scott, J. (2000). *Social Network Analysis: A Handbook*. Sage Publications.

Seifzadeh, S., Farahat, A. K., Kamel, M. S., & Karray, F. (2015). Short-text clustering using statistical semantics. In *Proceedings of the 24th International Conference on World Wide Web* WWW '15 Companion (pp. 805–810). New York, NY, USA: ACM.

Shehata, S., Karray, F., & Kamel, M. (2010). An efficient concept-based mining model for enhancing text clustering. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 1360–1371.

Shiau, W.-L., Dwivedi, Y. K., & Lai, H.-H. (2018). Examining the core knowledge on facebook. *International Journal of Information Management*, *43*.

Shiau, W.-L., Dwivedi, Y. K., & Yang, H. S. (2017). Co-citation and cluster analyses of extant literature on social networks. *International Journal of Information Management*, *37*, 390–399.

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*, *(in press)*.

Silver, D. L., Yang, Q., & Li, L. (2013). Lifelong machine learning systems: Beyond learning algorithms. In *AAAI Spring Symposium: Lifelong Machine Learning* (pp. 49–55).

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*, (pp. 1–21). URL: `https://doi.org/10.1007/s10479-017-2522-3`.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y., Saumya, S., & Roy, P. (2016). Predicting the helpfulness of online consumer reviews. *Journal of Business Research*, *70*.

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user metadata. *PLoS ONE*, *10*, e0115545. URL: `https://doi.org/10.1371/journal.pone.0115545`.

Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern Analysis and Applications*, *8*, 199–209.

Sowa, J. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Addison–Wesley.

Steinbach, M., Ertoz, L., & Kumar, V. (2003). The challenges of clustering high-dimensional data. In L. Wille (Ed.), *New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition* (pp. 273–307). Springer-Verlag.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168.

Subramaniam, L., Faruquie, T. A., Ikbal, S., Godbole, S., & Mohania, M. K. (2009). Business intelligence from voice of customer. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE 2009)* (pp. 1391–1402).

Sun, B., & Shibo, L. (2011). Learning and acting on customer information: A simulation-based demonstration on service allocations with offshore centers. *Journal of Marketing Research*, *48*, 72–86.

Thrun, S. (1998). Lifelong learning algorithms. In S. Thrun, & L. Pratt (Eds.), *Learning to learn* (pp. 181–209). Kluwer Academic Publishers.

Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. In *Proceedings of the 1st international conference on multidisciplinary information sciences & technologies* (pp. 354–358). volume II.

Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, *39*, 4729–4739.

Uysal, A., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*, 104–112.

Valbé, J.-J., Martí, M., Casanovas, P., Jakulin, A., Mladenic, D., & Fortuna, B. (2007). Stemming and lemmatisation: Improving knowledge management through language processing techniques. In P. Casanovas, P. Noriega, D. Bourcier, & F. Galindo (Eds.), *Trends on Legal Knowledge: the Semantic Web and the Regulation of Electronic Social Systems*. European Press Academic Publishing.

Vega-Vazquez, M., Revilla-Camacho, M. Á., & Cossío-Silva, F. J. (2013). The value co-creation process as a determinant of customer satisfaction. *Management Decision*, *51*, 1945–1953.

Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research*, *261*, 626–639.

Voorhees, C. M., Fombelle, P. W., Gregoire, Y., Bone, S., Gustafsson, A., Sousa, R., & Walkowiak, T. (2017). Service encounters, experiences and the customer journey: defining the field and a call to expand our lens. *Journal of Business Research*, *79*, 269–280.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge university press.

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3506–3510).

Yan, R., Cao, X.-b., & Li, K. (2009). Dynamic assembly classification algorithm for short text. *Acta Electronica Sinica*, *37*, 1019–1024.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 1445–1456).

Yan, X., Guo, J., Liu, S., Cheng, X.-q., & Wang, Y. (2012). Clustering short text using ncut-weighted non-negative matrix factorization. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2259–2262).

Yang, J., McAuley, J., & Leskovec, J. (2013). Community detection in networks with node attributes. In H. Xiong, G. Karypis, B. Thuraisingham, D. Cook, & X. Wu (Eds.), *Proceeding of the 2013 IEEE 13th International Conference on Data Mining* (pp. 1151–1156).

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 233–242).

Zhou, F., Zhang, F., & Yang, B. (2010). Graph-based text representation model and its realization. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering*. URL: `https://doi.org/10.1109/NLPKE.2010.5587861`.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, *48*, 379–398.