WHAT USERS WANT: A NATURAL LANGUAGE PROCESSING APPROACH TO DISCOVER USERS' NEEDS FROM ONLINE REVIEWS

Spada, Irene (1,5);
Barandoni, Simone (2,5);
Giordano, Vito (1,5);
Chiarello, Filippo (3,5);
Fantoni, Gualtiero (4,5);
Martini, Antonella (3,5)

1: School of Engineering, Department of Information Engineering, University of Pisa, Italy;
2: Department of Computer Science, University of Pisa, Italy;
3: School of Engineering, Department of Energy, Systems, Land and Construction Engineering, University of Pisa, Italy;
4: School of Engineering, Department of Civil and Industrial Engineering, University of Pisa, Italy;
5: B4DS - Business Engineering for Data Science lab, University of Pisa, Italy

ABSTRACT

Digital media are a means to deliver products and services, but also a channel to interact with consumers and a source of information on users' preferences. Data shared by customers on the web, the User-Generated Content (UGC), can give entrepreneurs a detailed perspective of the market. This work examines an application of Natural Language Processing techniques on UGC to discover insights on users' opinions. We collected more than 13.000 reviews of software from digital stores and review website to gather information on the customers' perspective and their response to a given marketing strategy in two case studies on digital product's launch. The objective is to give support to two Italian companies in the process of business model development through data-driven evidence. We aim to discover who are the users and which are their needs using a lexicon-based approach to mine unstructured text. The results provide qualitative and quantitative descriptions of the market segments. We propose a method to examine UGC and to explore customers' behavior on social media. The findings helped managers for the development of their business model, enhancing an informed decision-making process.

Keywords: Semantic data processing, User centred design, Requirements

Contact:
Spada, Irene
University of Pisa
Italy
irene.spada@phd.unipi.it

Cite this article: Spada, I., Barandoni, S., Giordano, V., Chiarello, F., Fantoni, G., Martini, A. (2023) 'What Users Want: A Natural Language Processing Approach to Discover Users' Needs from Online Reviews', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.389

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

1 INTRODUCTION

Digital transformation is changing the competitive dynamics and is calling for new strategies and models to handle emerging challenges. The reduction of product life cycles and the faster introduction of new ones, requires up to date information to fulfil customers' needs (Chiarello et al., 2021). This evolution is also producing new opportunities related to the exploitation of digital media in business. Digital media can be intended not only as a means to deliver products and services, but also as a channel to interact with consumers, and as a source of information on users' preferences (Vlačić et al., 2021). Indeed, data posted and shared by customers on social media and web-platforms, the so-called User-Generated Content (UGC), are a key resource for gathering the opinions of the users (Tang et al. 2022) and give to the entrepreneurs a detailed perspective of the market segment (Checchinato, 2021). Analysing these documents through automatic ways allows to significantly reduce the time commonly needed to read them. Therefore, timely outcomes might be produced by exploiting technology to collect knowledge from large amount of UGC. Many authors propose several Natural Language Processing (NLP) approaches to mine UGC and discover needs, product features, and their links. Notwithstanding, few studies focus on the use of online content also for the identification of the users.

In the present work, we propose the application of NLP techniques to mine reviews and posts of digital users and gather insights on the customers' perspective and their response to a given marketing strategy, leveraging on the description of users and needs available in the UGC. Specifically, we aimed to enlarge the perspective in the analysis of UGC to discover both who were the target customers, and which were their needs, answering to the following Research Question:

RQ*: How do customers describe themselves and their needs in online reviews?*

Answering to this question by exploiting automatic NLP techniques can provide valuable knowledge in short time on current market situation and users' interests, useful to support the processes of design, development, production, and commercialization of a product. We performed a data-driven market analysis using a dataset of reviews from different web-platforms in two case studies on digital product's launch. Two small companies in Italy were developing online applications and were interested in understanding the profitability of their product. Our objective was to give them support in the process of economic feasibility and business model development through data-driven evidence. The empirical setting (digital product) is characterised by intangible products with zero marginal cost and stored on digital media and communities of customers acting on digital environments (Öberg and Alexander, 2019). These features make competition in digital markets different rather than physical markers and the use of UGC fundamental to develop and implement business strategies (Saura et al., 2021). Our analysis led to the identification of the customer segments to include in the target market and the most relevant requirements that the applications should satisfy. The gathered insights, in the analysed cases, allowed us to support the two companies in the process of decision making during the development of their business model. In general, our study can benefit both researchers and practitioners. The proposed approach can effectively detect users and needs in UGG, and it can be applied to different data sources and in various domains to explore customers' behaviour. The elicitation of users and needs from UGC can provide timely and updated information that can be useful for companies in the various phases of product development, from the definition of product design specifications to the delineation of the marketing strategy. This regards not only the domain of software design and development, but also the engineering design and the production of physical products: knowledge extracted from UGC might be valuable for the design and commercialization processes of a physical system as well.

2 BACKGROUND LITERATURE

New opportunities for text analysis derive from data posted and shared by the customers in social media or other web-platforms, the so-called User-Generated Content. Scholars and practitioners are experimenting new ways of using big data in forecasting research, and the UGC analysis lies in this stream of literature. This kind of analysis deals with processing and examining data coming from social media, which are a rich source of information for the preferences and the opinions of the users. UGC data include text and images and come from various online resources, such as social media, shopping websites, blogs and in general every website where online users can leave a comment (Tang et al., 2022). Those resources are a valuable source of information for customers' needs, which are the

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

benefits that customers want to achieve by using a given product or service, typically expressed in natural language (Griffin and Hauser, 1993).

2.1 Content of UGC

Several authors exploit data shared by users and Artificial Intelligence techniques to delineate customers' needs and support business strategies. Kühl et al. (2020) use a supervised machine learning algorithm on tweets to classify customers' needs. Others propose evaluating reviews to delineate the rationale behind decisions (Kurtanović and Maalej, 2017) or identify customer behaviour patterns (De Luca et al., 2021). Customers' reviews can be used also to elicit products or services characteristics. Zhang et al. (2018) present a framework for using online product purchase data to characterise the features of a given product. Ding et al. (2008) propose a lexicon-based approach to mine UGC, starting from a previous work in literature they enlarged a list of terms and expressions typically used to communicate opinions, then they use this list to identify the orientation of the comments on product features, i.e., to assign a sentiment to the reviews. Abrahams et al. (2015) use an automotive-related lexicon to mine UGC and identify components of vehicles and then their defects. Customers' needs and products or services characteristics are related in the context of Engineering Design to frame the scope and guide the development process (Chiarello et al., 2021). Moreover, they can sometimes overlap in UGC: Harding et al. (2001) state that customers may express their needs also referring to technical requirements. Indeed, many scholars address these connections. Johann et al. (2017) present a rule-based method to extract the features of a given product to contrast and compare the items identified in the app descriptions of the applications written by developers and the reviews of the users. A tree-structured classification algorithm is used to map customer preferences for customised product design (Du et al., 2003). Besides, Wang et al. (2020) proposes a deep learning-based approach to link customers' needs and products' features using reviews and lists of design specifications. Finally, Li et al. (2022) propose a BERT model in combination with a Bayesian Network to extract and evaluate product attributes aiming at understanding the potential factors which affect the customers' purchase preferences.

2.2 Helpfulness and reliability of UGC

User Generated Content can provide insights into the actual data of customers and people in general, without any modification and/or filter of regular media (Krumm et al., 2008). Those can be useful both for customers, to guide their preferences, and for companies, to understand their target market. Many researchers discuss the utility of the reviews. Pan and Zhang (2011) analysed the combined effects of reviews' characteristics and product types (i.e., experiential vs. utilitarian products) to measure the perceived helpfulness. Wang et al., (2011) exploited UGC to estimate customers' preferences and created models for the design selection of a physical product (considering a smartphone as an example). Park et al., (2023) proposed a neural network model to automatically analyse online data to identify smartphone specification configurations preferred by customers and to support companies in the design process. Recently, researchers elaborated on the warranting theory to explain the trust of people in online reviews, highlighting that people are more likely to choose and evaluate positively a given product or service when they are confident about the fact that actual customers share the online reviews (DeAndrea et al., 2018). This opens the discussion to the reliability of the UGC. Indeed, the use of online reviews can raise the concern about their truthfulness: as anyone can post a review, it is easy to stumble upon fraudulent comments, the so-called fake reviews, whose purpose is to promote or devalue products and services; they can be generated by real persons, hired for this task, or even by robots, which can automatically insert comments online (Kauffmann et al., 2020). Some scholars consider the posting frequency of the users for fake detection (Jain et al., 2021). Others consider textual features and metadata of reviews (Fontanarava et al., 2017). Several NLP techniques have been adopted for this purpose, such as lexical, syntactic, and semantic analysis (Kauffmann et al., 2020), as well as a quantitative validation procedure for ensuring the reliability of online reviews (Barravecchia et al., 2021). Finally, the credibility of the source plays an important role for the accuracy of the reviews and their acceptance by online users (Kiecker and Cowles, 2002; O'Reilly et al., 2016). In conclusion, the comments/reviews, posts, and blogs can be mined to obtain interesting and valuable insights on the customers' perspective and response to a given marketing strategy. The content of the reviews can include both customers' needs and products' features. That information can support customers and companies in the decision-making process, especially when their trustworthiness is ensured. Despite the large literature on this topic, there is a lack of studies on the identification of the users. Indeed, all the cited works focus more on needs,

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

product features, and their comparison, and they provide a characterization of the users based on their needs. Our proposed paper aims at leveraging on NLP and enlarging the perspective, considering both the who (users) and the what (needs).

3 DATA AND METHOD

3.1 Case studies

We examined customers' reviews in two case studies related to digital products, described below (*the companies are indicated with fictional names for the purpose of privacy*).

- **Case Study 1**: *GemData* is a small company in Tuscany dedicated to applications and website development for businesses. The customers in their portfolio are located principally at the local level, they also have some contacts at the national level. They plan to launch an online data-sharing platform and enter the market of cloud solutions for small and medium-sized enterprises.
- **Case Study 2**: *SplendidTech* is a small company in Tuscany operating in software and IT infrastructures development, integration and maintenance. They operate mainly at the local and national level, but they also have European customers in their portfolio. They aim to introduce a customizable application for remote collaboration and for distance learning in the market.

We decided to select these two cases because the work done for the two companies offered us the opportunity to deepen the domains of online data-sharing platforms and collaboration software. Digital market is already full of software with similar characteristics to these ones. The intense competition makes their commercialization harder. At the same time, the digital markets provide data coming from customers' communities (Öberg and Alexander, 2019), allowing to retrieve information useful for the analysis (e.g., competitors' software features and user reviews). Next, the work of these two companies can represent an exemplary case for the analysis that can be conducted to support the commercialization of a product of any domain which is going to be launched in a digital marketplace or web store.

3.2 Data collection

Our purpose was to retrieve users and needs through a data-driven approach. We decided to utilise software reviews as a source of User-Generated Content in the data collection phase. Reviews are a rich source of information for customer segments and needs. Users tend to communicate their profession to better explain the reasons behind the choice and the use of a given product or service. They usually evaluate products or services relying on their experience with it. They also mention some features they appreciate, or elements not working as they would expect, or missing characteristics. These elements represent useful information in the identification of users and needs. Initially, we determined the relevant software category for each of the presented cases. We searched for other products with similar technical characteristics on digital stores and software review websites. The accessed resources were some general-purpose platforms, used to browse and download applications and games for many devices, and specific online platforms, used to rate and compare software. We utilised these platforms to retrieve users' comments. We collected more than 13.000 reviews of software. We obtained this data from two digital stores and one software review website. We selected the sources that best fit the objectives of the analysis in each of the two cases: we utilised Apple App Store, Google Play Store and Capterra.com for the collection of reviews. We relied on two lexicons to retrieve the relevant information from the reviews: a list composed of 10.219 target-groups built on an extraction algorithm applied on patents (Chiarello et al., 2018) and the list of standard occupations included in the European Skills, Competences, Qualifications and Occupations (ESCO) framework were used to detect user categories. We applied a list of sixty-four system quality attributes (Adams, 2015) for the customer needs. Those attributes describe the architecture of a software, and so the scope of a given software, indeed, as reported by Harding et al. (2001), customers' needs can be also formulated as system quality attributes of a product or service.

3.3 Data processing: Users and needs identification

We prepared the data for the analysis. First, we assessed the truthfulness of the reviews. For Case Study 1, the text statistics indicators derived from Li et al. (2011) and Fontanarava et al. (2017). For Case Study 2, we did not apply the same methodology to assess the truthfulness of the reviews because in this case the textual data were collected entirely from a software review website (i.e.,

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

Capterra), where reviews are written in a well-structured and reliable way. We referred to the studies of Kiecker and Cowles (2002) and O'Reilly et al. (2016), which state that the reliability of the source can ensure the truthfulness for reviews. We applied standard cleaning operations to all textual data: lowercasing, symbols and numbers removal, missing values (such as empty reviews) deletion. We carried out a further pre-processing operation on the list of System Quality Attributes. An examination of a random sample of the customers' review allowed us to understand that customers are likely to express needs also through the corresponding adjectives. For instance, *accessible* can convey the same meaning of *accessibility*. For this reason, during pre-processing, we also applied the stemming procedure on all the System Quality Attributes to have terms in base form. The stems do not include any suffixes, leading us to detect needs when expressed both in the adjective and the noun forms. The methodology adopted to extract user's needs and user categories was a gazzetter-based approach, where the lexicons explained in section 3.2 were the gazetteers used in a string searching algorithm. It compares strings, and can be considered a white box model, able to match the occurrences of System Quality Attributes and user categories in the texts only if present there. We counted the occurrences of each extracted term of the different lexicons in the two datasets of reviews for each case study. Then, we computed their relative frequency, i.e., the occurrences divided by the number of the reviews that contain at least one user or one need. These measures are widely used to represent the importance of a given term in a text. In our case, these metrics represent the relevance of the user or of the need within the context of the selected digital platform.

4 RESULTS

4.1 Results of data collection

We collected the reviews of software from digital stores and one software review website. As anticipated, we used different sources in relation to the main objectives of the analysis in the two cases. Both companies were interested in timely data on the target market to accurately plan the launch of the new products. *GemData* was interested in delineating the size and the needs of the target customer segments to ensure alignment with market demand. *SplendidTech* was concerned with understanding the needs of the target customers in the sector. We created two datasets, one for each of the two case studies. For Case Study 1, we collected 23.468 reviews from digital stores, namely Google Play Store and App Store, where users can also rate an app without giving explanations. These platforms allow us to keep a broad perspective on the market. For Case Study 2, we retrieved 1.250 reviews from a software review website, namely Capterra, where users are requested to write much longer and complex reviews compared to the ones in digital stores. This choice lets us deepen the characteristics of the sector. The analysis of similar applications allowed us to identify the relevant software category for each case. The online platform for data sharing by *GemData* belongs to the cloud storage application category, which are repositories accessible on the Internet from any device or location for archiving, maintaining, and retrieving data and documents online; the Internet connection allows files to be synchronised and updated. This includes among others Google Drive; Bitrix24; Canto DAM; Kamzam; Degoo; Pcloud; Tresorit; Thron; SendAnywhere. Online collaboration platform category, namely software that combines communication tools, such as messaging and video calls, sharing and collaboration functions in a virtual workspace, and tools to manage workflows remotely, encompasses the customizable applications for remote collaboration/learning by *SplendidTech*. This category includes Airtable; Amazon Chime; ClickUp; Dropbox Business; Flock; Google Workspace; Microsoft Teams; Quire; Trello.

4.2 Results of data processing

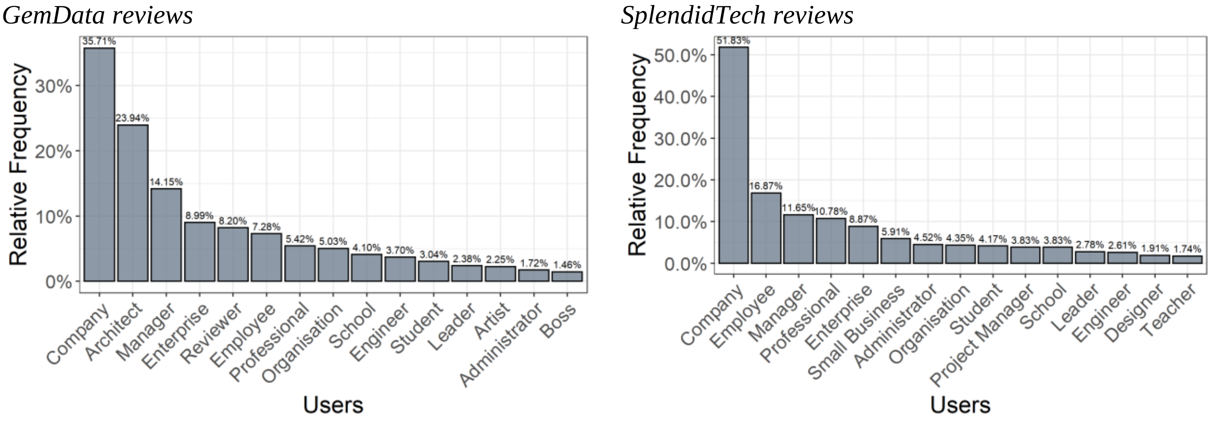
The pre-processing procedure leads to removing fake and useless reviews from the raw dataset and cleaning it to have more reliable information (Kauffmann et al., 2020). The number of possible fake reviews in Case Study 1 is 4.223, which is 18% above the total. We manually checked some phrases, but we cannot be certain if they are fake reviews or not. However, they have no content that could be relevant to our analysis, for example "*Omg! Love this app so much!*", "*nice app if you like free storage on cloud*"; or even "*Hi, As the app name and description says, Degoo is a cloud drive app. It will give you 100 GB of online storage space where you can upload your files. It will not increase your phone's storage space or give you free mobile data. If you need more information for using Degoo, please see*

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

this guide (https://goo.gl/Jbku8L).", which has been found many times with different users' names. For Case Study 2 the reliability of the source can ensure the truthfulness for reviews. The Terms of Use of Capterra (2022) requires "sufficient information for our Quality Assurance team to verify their identity" and "[we] remove reviews that we in our discretion determine do not comply with our Community Guidelines". Therefore, the collection of reviews can be considered reliable. The procedures of text cleaning and fake removal led to 12.000 reviews for Case Study 1, while for Case Study 2 we kept all the 1.250 reviews. We specify that the total number of reviews removed in Case Study 1 (namely, 11.468), although very high, is reasonable due to the nature of the Play Store and App Store reviews, where users can leave comments even without text and only by assigning a star rating. On average, the number of words per review is 38 for App Store and Play Store (Case Study *GemData*), and 226 for Capterra (Case Study *SplendidTech*). The string-matching algorithm allowed us to identify respectively 1.054 and 892 users and 4.895 and 2.307 needs for *GemData* and *SplendidTech*. Considering the unique values, we detected 37 and 71 users and 45 and 54 needs for *GemData* and *SplendidTech*. At least one user has been found in 6,3% of reviews for *GemData* and in 46% of reviews for *SplendidTech*; and at least one need has been found in 31% of reviews for *GemData* and 76% for *SplendidTech*.

4.2.1 Customer segments

The data extracted from the reviews using the lexicon of target-groups and the list of ESCO occupations gives us some indication about the types of users potentially interested in using applications like those proposed by *GemData* and *SplendidTech*. The top 15 types of users are presented in the charts in Figure 1, respectively for *GemData* and *SplendidTech*. The relevance of the types of users is proportional to their relative frequency, that is the number of different reviews in which a given type of user is mentioned divided by the total number of reviews in which at least one type of user is mentioned.


GemData reviews

SplendidTech reviews

Figure 1. Barplots of the 15 most relevant types of users extracted from reviews.

The first position in both charts in Figure 1 (i.e., company) is very general and not meaningful for the purpose of our study. However, customers tend to write *I work in a small (or big) company* to indicate their job. The description of their profession is sometimes indicated with the location: in the charts it is possible to find *school* or *enterprise*, for instance. In some cases, users specify the size of the company, e.g., *small business*. Many times, customers communicate their position in the company, e.g., *employee*, *manager*, or *leader*. Some indicate their specific occupation, such as *architect*, *engineer*, *designer*, or even *student*. The identified users can be grouped in several categories to delineate the customer segments. Table 1 presents some examples of the detected users divided per segment in the case studies. The most relevant customer segments for *GemData* are, among others, computer scientists, musicians, and photographers; in fact, they need cloud spaces to store and share data. The product proposed by *SplendidTech* interested mostly project managers, business consultants, and HR staff, as they need collaborative platforms for team working. Furthermore, both applications are used by engineers, professors, and students. The full lists of possible users were discussed in a dedicated brainstorming session with managers of the two companies, aiming at identifying the segments to include in the target. *GemData* is a newcomer in the market of cloud storage applications. Its management decided to not target medium-large companies and/or organisations such as schools and hospitals, which are typically

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

not included in their client portfolio. These types of companies are likely to already have a cloud storage and sharing system, or they probably prefer to rely on larger and already established competitors. The defined target market is therefore the set of professionals in small businesses and freelancers from the technical sectors (architects, engineers, surveyors, computer scientists), legal (lawyers) and artistic/cultural (photographers and musicians). The application proposed by *SplendidTech* has already been tested with associations in the field of training and remote education. The company does not intend at this stage to target medium-large companies and/or organisations operating in the healthcare sector or in other sectors. Therefore, Education, Healthcare and Others have not been included in the target market. The target market encompasses the group of Business and Professionals, addressing both SMEs and medium-large enterprises. In addition, users of particular interest were identified for each category. The list of selected users includes project manager; business consultant; salesperson; project management group; HR function; engineer; designer; attorney; legal; paralegal.

Table 1. Types of users extracted from the reviews, organised in customer segments.

Case Study	**Users identified in the reviews (examples)**	**Customer Segments**
GemData	Engineers, Computer Scientists, Architects, Programmers, Lawyers, Musicians, Photographers	Professionals and freelances
	Professors, Teachers, Students, Technical Staff	Education
	Hospital Administration, Clinics	Healthcare
SpendidTech	Engineers, Managers, Consultants, Programmers, Web designers, System Developers, Lawyers, Musicians, Photographers	Business and Professionals
	Professors, Teachers, Students, Technical Staff	Education
	Hospital, Clinics	Healthcare
	Volunteer, Ambassador, Non-profit	Others

4.2.2 Customers' needs

The needs of the customers interested in the application proposed by *GemData* and *SpendidTech* have been detected from the review using the list of sixty-four System Quality Attributes (Adams, 2015). Figures 2 reports the charts of the top 10 needs respectively for *GemData* and *SpendidTech*. The relevance of the needs is based on their relative frequency, that is the percentage of reviews in which a given need is cited divided by the total number of reviews in which at least one need is mentioned.

We observe that the most important features for the customers are those which allow a product to be simple to use, secure, and easy to access. Customers express the need to use the application also on mobile devices for both digital products. The needs of customer segments for *GemData* also include safety: cloud platforms must ensure the protection of the data in uploading, storing, and sharing operations. Besides other needs, the interested customer segments for *SplendidTech* also look for flexibility. Indeed, due to the increasing use of collaborative platforms, users would like to have more and more features to interact with on digital spaces. The obtained results helped managers of both *GemData* and *SpendidTech* in the process of business model development. The quantitative insights demonstrate which are the most important needs for the target users. The proposed applications should therefore include technical and specific features to satisfy the requirements. The management team and the developers of *GemData* and *SpendidTech* used this information to estimate if their software's technical characteristics were enough to fulfil the market's requirements or if they needed a further development to improve some aspects of their products.

The distinctive features of the online platform for data sharing by *GemData* are the following: internal messaging system, rapid search bar for documents, possibility to create groups, possibility to upload documents in different formats, mobile accessibility, user-friendly design, privacy and security rules for users and documents. These features are in line with the main needs of users identified in the analysis and presented in the previous sections: the simplicity of use and the security of the platform, the ease of access to documents by different users and devices and the overall reliability of the platform.

The characteristics of the applications for remote collaboration/learning by *SplendidTech* are the following: tools to design and schedule workshops, events, and brainstorming sessions, access controls/permissions, file sharing, tutorials, customization of the interface. Those features satisfy the needs of the target users, discussed in the previous sections, such as the possibility of customization and the flexibility of the working environment for different operative conditions, the simplicity of use

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

of the platform and the security of the platform, since the platform is designed in compliance with the highest security standards and in compliance with European data privacy and security standards.

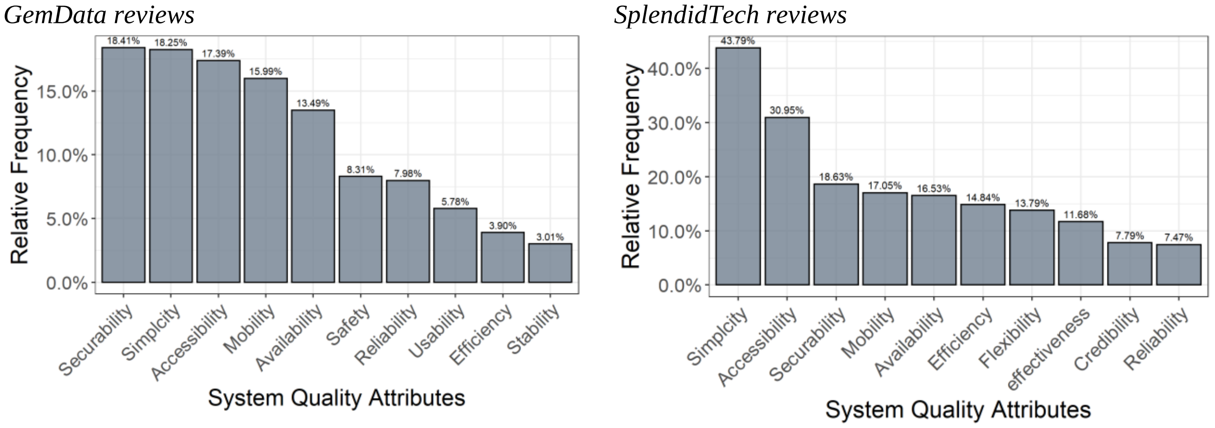


Figure 2. Barplots of the 10 most relevant system quality attributes extracted from reviews.

5 CONCLUSION

This work examines an application of NLP techniques on UGC to discover insights on users' opinions. We rely on more than 13.000 reviews from different platforms to support the marketing strategy development in two case studies on digital product’s launch. The proposed case studies demonstrate that automatic analysis of reviews can effectively support decision making processes. From a managerial perspective, a method for identifying key information on user preferences can benefit practitioners in assessing the relevance of users and needs for a given product. Indeed, the results of the analysis provide qualitative and quantitative descriptions of the market segment in which a certain product can be distributed. The application of NLP on reviews of users on digital platforms can provide data-driven insights on customers and so enhance an informed decision-making process in the phases of product development, from product specifications’ design to marketing strategy definition. On one hand, the needs identification analysis can help developers to understand if their software fulfils the customer demands or if they need further work to add or improve a particular feature before the commercialization process. On the other hand, the customer segments analysis can give companies an overview on the types of users that are going to be interested in their product. This is valuable information that can be utilised both to sharpen a certain software to make it more attractive for those categories of people and to design an effective and targeted advertising campaign. Also, designers of physical systems could benefit from the application of the proposed methodology as well. Nowadays, physical products are more and more embedded in software applications. As we know this can make the design processes more complex and the need of data-driven decision making in more evident. Therefore, our method is a step in this direction positioned in the stream of literature of data driven design (Chiarello et al., 2021).

From an academic perspective, the proposed approach is an effective method to examine UGC. This work is in line with most of the studies in the related stream of literature. Furthermore, it provides a relevant contribution to the stream of user characterization from reviews. We found out that most of the previous literature revolved around the content of the reviews to detect needs and features. Few spaces are given to user identification. So, we focus on how users tend to describe themselves while writing about a product. We demonstrated that the approach can be easily replicated with different data sources. Therefore, researchers may be interested to apply this kind of approach in different platforms and/or social media, or even in different domains, to explore differences in customers’ behaviour. Furthermore, even though we considered two case studies concerning the development of software product, the proposed methodology might also be applied to enhance the commercialization process a physical product. This would require the identification of adequate data sources, e.g., review websites dedicated to a specific physical product, the definition of that product’s specific technical characteristics and the relation between these features and the users’ needs. Our proposed work has some limitations, which could be addressed in future research works. Our approach relies on a gazetteer-based Named Entity Recognition method (Pawar et al., 2012), i.e., we use a list of terms to identify relevant elements in unstructured text. This kind of approach leads to high precision since the values derive from a validated list. In contrast, it brings low values in recall, as the completeness of the

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

lexicon influences the number of extracted entities. The combination of different approaches, like rule-based approach and machine learning algorithms can improve the quality of the results. Another limitation concerns the lack of a structured model to link customer needs, retrieved through System Quality Attributes, to product requirements, to understand directly the potentially missing features. A future improvement could be integrating NLP techniques with design tools, such as Quality Function Deployment (QFD) used to link user needs to technical requirements, to produce a data-driven model for better support design processes.

ACKNOWLEDGEMENTS

This study was founded by a regional project *Bando per l'assegnazione di voucher per l'innovazione alle MPMI* (Call for assignment of innovation vouchers in SMEs; approval *Giunta Camerale* n.12 - 28 Jan. 2021; CUP code: G49J21001180005; COR code: 5152003). The project was promoted by *Camera di Commercio della Maremma e del Tirreno* and realised by University of Pisa and Beam me up, a spin-off of University of Pisa. We thank Beam me up (https://www.beammeup.it/) and Costanza Soffici for her precious work in interacting with the companies.

REFERENCES

Abrahams, A.S., Fan, W., Wang, G.A., Zhang, Z. and Jiao, J., (2015), "An integrated text analytic framework for product defect discovery.", *Production and Operations Management*, Vol. 24 No. 6, pp. 975-990. https://doi.org/10.1111/poms.12303

Adams, K.M., (2015), *Nonfunctional requirements in systems analysis and design*, Cham: Springer international publishing. https://doi.org/10.1007/978-3-319-18344-2_3

Barravecchia, F., Mastrogiacomo, L., and Franceschini, F. (2021), "Digital voice-of-customer processing by topic modelling algorithms: insights to validate empirical results.", *International Journal of Quality & Reliability Management, Vol. 39 No. 6, pp.1453-1470.* https://doi.org/10.1108/IJQRM-07-2021-0217

Capterra (2022, January), *Capterra Terms of Use, r*etrieved February 2023, from Capterra, Capterra General User Terms. Available at: https://www.capterra.com/legal/terms-of-use

Checchinato, F. (2021), "Digital transformation and consumer behaviour: How the analysis of consumer data reshapes the marketing approach.", In: Hinterhuber, A., Vescovi, T., & Checchinato, F. (Eds.), *Managing Digital Transformation: Understanding the Strategic Process*., Routledge, London, pp. 165-176. https://doi.org/10.4324/9781003008637

Chiarello, F., Cimino, A., Fantoni, G., and Dell'Orletta, F. (2018), "Automatic users extraction from patents.", *World Patent Information*, Vol. 54, pp. 28-38. https://doi.org/10.1016/j.wpi.2018.07.006

Chiarello, F., Belingheri, P., and Fantoni, G. (2021), "Data science for engineering design: State of the art and future directions.", *Computers in Industry*, Vol. 129, p. 103447. https://doi.org/10.1016/ j.compind.2021.103447

DeAndrea, D. C., Van Der Heide, B., Vendemia, M. A., and Vang, M. H. (2018), "How people evaluate online reviews.", *Communication Research*, Vol. 45 No. 5, pp. 719-736. https://doi.org/ 10.1177/0093650215573862

De Luca, L. M., Herhausen, D., Troilo, G., and Rossi, A. (2021), "How and when do big data investments pay off? The role of marketing affordances and service innovation.", *Journal of the Academy of Marketing Science*, Vol. 49 No. 4, pp. 790-810. https://doi.org/10.1007/s11747-020-00739-x

Ding, X., Liu, B., and Yu, P. S. (2008), "A holistic lexicon-based approach to opinion mining.", *Proceedings of the 2008 international conference on web search and data mining*, pp. 231-240. https://doi.org/ 10.1145/1341531.1341561

Du, X., Jiao, J., and Tseng, M. M. (2003), "Identifying customer need patterns for customization and personalization.", *Integrated manufacturing systems*, Vol. 14 No. 5, pp. 387-396. https://doi.org/10.1108/ 09576060310477799

Fontanarava, J., Pasi, G. and Viviani, M. (2017), "Feature analysis for fake review detection through supervised classification.", 2017 *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Tokyo, Japan, pp. 658-666. https://doi.org/10.1109/DSAA.2017.51

Harding, J. A., Popplewell, K., Fung, R. Y., and Omar, A. R. (2001), "An intelligent information framework relating customer requirements and product characteristics.", *Computers in Industry*, Vol. 44 No. 1, pp. 51-65. https://doi.org/10.1016/S0166-3615(00)00074-9

Jain, P. K., Pamula, R., and Ansari, S. (2021), "A supervised machine learning approach for the credibility assessment of user-generated content.", *Wireless Personal Communications*, Vol. 118 No. 4, pp. 2469-2485. https://doi.org/10.1007/s11277-021-08136-5

Johann, T., Stanik, C., and Maalej, W. (2017), "Safe: A simple approach for feature extraction from app descriptions and app reviews.", 2017 *IEEE 25th international requirements engineering conference (RE)*, Lisbon, Portugal, 2017, pp. 21-30. https://doi.org/10.1109/RE.2017.71

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., and Mora, H. (2020), "A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making.", *Industrial Marketing Management*, Vol. 90, pp. 523-537. https://doi.org/10.1016/j.indmarman.2019.08.003

Kiecker, P., and Cowles, D. (2002), "Interpersonal communication and personal influence on the Internet: A framework for examining online word-of-mouth.", *Journal of Euromarketing*, Vol. 11 No. 2, pp. 71-88. https://doi.org/10.1300/J037v11n02_04

Krumm, J., Davies, N., and Narayanaswami, C. (2008), "User-generated content.", *IEEE Pervasive Computing*, Vol. 7 No. 4, pp. 10-11. https://doi.org/10.1109/MPRV.2008.85

Kühl, N., Mühlthaler, M., and Goutier, M. (2020), "Supporting customer-oriented marketing with artificial intelligence: automatically quantifying customer needs from social media.", *Electronic Markets*, Vol. 30 No. 2, pp. 351-367. https://doi.org/10.1007/s12525-019-00351-0

Kurtanović, Z., and Maalej, W. (2017), "Mining user rationale from software reviews.", 2017 *IEEE 25th international requirements engineering conference (RE)*, Lisbon, Portugal, 2017, pp. 61-70. https://doi.org/10.1109/RE.2017.86

Li, F. H., Huang, M., Yang, Y., and Zhu, X. (2011), "Learning to identify review spam.", *Twenty-second international joint conference on artificial intelligence*, Barcelona, Spain, 2011. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-414

Li, M. F., Zhang, G. X., Zhao, L. T., and Song, T. (2022), "Extracting product competitiveness through user-generated content: A hybrid probabilistic inference model.", *Journal of King Saud University-Computer and Information Sciences*, Vol 34 No. 6, pp. 2720-2732. https://doi.org/10.1016/j.jksuci.2022.03.018

Öberg, C., and Alexander, A. T. (2019), "The openness of open innovation in ecosystems–Integrating innovation and management literature on knowledge linkages.", *Journal of Innovation & Knowledge*, Vol. 4 No. 4, pp. 211-218. https://doi.org/10.1016/j.jik.2017.10.005

O'Reilly, K., MacMillan, A., Mumuni, A. G., and Lancendorfer, K. M. (2016), "Extending our understanding of eWOM impact: The role of source credibility and message relevance.", *Journal of Internet Commerce*, Vol. 15 No. 2, pp. 77-96. https://doi.org/10.1080/15332861.2016.1143215

Pan, Y., and Zhang, J. Q. (2011), "Born unequal: a study of the helpfulness of user-generated product reviews.", *Journal of retailing*, Vol. 87 No. 4, pp. 598-612. https://doi.org/10.1016/j.jretai.2011.05.002

Park, S., Joung, J., and Kim, H. (2023), "Spec guidance for engineering design based on data mining and neural networks.", *Computers in Industry*, Vol. 144, p. 103790. https://doi.org/10.1016/j.compind.2022.103790

Pawar, S., Srivastava, R., and Palshikar, G. K. (2012, January), "Automatic gazette creation for named entity recognition and application to resume processing.", *Proceedings of the 5th ACM COMPUTE Conference: Intelligent & scalable system technologies*, pp. 1-7. https://doi.org/10.1145/2459118.2459133

Saura, J. R., Ribeiro-Soriano, D., and Palacios-Marqués, D. (2021), "From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets.", *International Journal of Information Management*, Vol. 60, p. 102331. https://doi.org/10.1016/j.ijinfomgt.2021.102331

Tang, L., Li, J., Du, H., Li, L., Wu, J., and Wang, S. (2022), "Big Data in Forecasting Research: A Literature Review.", *Big Data Research*, Vol. 27, p. 100289. https://doi.org/10.1016/j.bdr.2021.100289

Vlačić, B., Corbo, L., e Silva, S. C., and Dabić, M. (2021), "The evolving role of artificial intelligence in marketing: A review and research agenda.", *Journal of Business Research*, Vol. 128, pp. 187-203. https://doi.org/10.1016/j.jbusres.2021.01.055

Wang, L., Youn, B. D., Azarm, S., and Kannan, P. K. (2011), "Customer-driven product design selection using web based user-generated content.", *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 54822, pp. 405-419. https://doi.org/10.1115/DETC2011-48338

Wang, Y., Luo, L., and Liu, H. (2020), "Bridging the semantic gap between customer needs and design specifications using user-generated content.", *IEEE Transactions on Engineering Management*, Vol. 69, pp. 1622 - 1634. https://doi.org/10.1109/TEM.2020.3021698

Zhang, J., Simeone, A., Gu, P., and Hong, B. (2018), "Product features characterization and customers' preferences prediction based on purchasing data.", CIRP Annals, Vol. 67 No. 1, pp. 149-152. https://doi.org/10.1016/j.cirp.2018.04.020

https://doi.org/10.1017/pds.2023.389 Published online by Cambridge University Press