Industrial Marketing Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Industrial Marketing Management

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

Research paper

A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making

Erick Kauffmann[a], Jesús Peral[b], David Gil[c,*], Antonio Ferrández[b], Ricardo Sellers[d], Higinio Mora[c]

[a] Department of Industrial Engineering, University of Costa Rica, San José, Costa Rica
[b] Department of Software and Computing Systems, University of Alicante, Alicante 03690, Spain
[c] Department of Computing Technology and Data Processing, University of Alicante, Alicante 03690, Spain
[d] Department of Marketing, University of Alicante, Alicante 03690, Spain

ARTICLE INFO

Keywords:
Big data analytics
Sentiment analysis
Marketing decisions
High-tech industries
Fake reviews

ABSTRACT

User-generated content about brands is an important source of big data that can be transformed into valuable information. A huge number of items are reviewed and rated by consumers on a daily basis, and managers have a keen interest in real-time monitoring of this information to improve decision-making. The main challenge is to mine reliable textual consumer opinions, and automatically use them to rate the best products or brands. We propose a framework to automatically analyse these reviews, transforming negative and positive user opinions in a quantitative score. Sentiment analysis was employed to analyse online reviews on Amazon. The Fake Review Detection Framework—FRDF— detects and removes fake reviews using Natural Language Processing technology. The FRDF was tested on reviews of products from high-tech industries. Brands were rated according to consumer sentiment. The findings demonstrate that brand managers and consumers would find this tool useful, in combination with the 5-Star score, for more comprehensive decision-making. For instance, the FRDF ranks the best products by price alongside their respective sentiment value and the 5-Star score.

1. Introduction

Big data represents one of the most important challenges for academics and practitioners. Every day, huge amounts of data from different sources are generated at great velocity. The ability of firms and managers to transform these data into valuable information could make the difference between business success and failure. In this sense, big data analytics is rapidly becoming a trending practice adopted by many organizations with the aim of obtaining valuable information (Sivarajah, Kamal, Irani, & Weerakkody, 2017). However, there is limited knowledge on how firms and organizations transform the potential offered by big data into real social and economic value (Günther, Mehrizi, Huysman, & Feldberg, 2017). Hence, there is growing demand for applications that can rapidly provide big data analytics for businesses (Lytras, Raghavan, & Damiani, 2017; Chen, Chiang, & Storey, 2012).

From a marketing perspective, the importance of big data highlights the need to understand how big data analytics could shape, sense and respond to real customer demands (Kumar, Shankar, & Aljohani, 2019). However, research on big data applications for marketing is still at embryonic stage, making it essential to increase efforts for big data to be acknowledged as a key tool in the marketing field (Amado, Cortez, Rita, & Moro, 2018).

Among the different sources of big data, User-Generated Content (UGC) is one of the most important ones. From blogs to social media and online reviews, consumers generate huge amounts of brand related information that have a decisive potential business value in targeted advertising (Zhang & Katona, 2012), customer-firm relationships (Chuang, 2019) or brand communication (De Vries, Gensler, & Leeflang, 2012), among others. In the same line, previous empirical findings show that UGC has significant effects on brand images, purchase intentions, and sales (De Vries et al., 2012).

One important type of UGC is electronic word-of-mouth (eWOM) expressed by users, as it plays an important role for customer potential buying decisions (Wang, Xu, Fujita, & Liu, 2016). As stated by Feldman (2013), the decision-making process of people is affected by the opinions formed not only by thought opinion leaders, but also from ordinary people. Consumers usually search for reviews and opinions written by other people when they want to buy a product online. Thus, mining and analysing UGC data such

* Corresponding author.
E-mail address: dgil@dtic.ua.es (D. Gil).

https://doi.org/10.1016/j.indmarman.2019.08.003
Received 15 November 2018; Received in revised form 27 July 2019; Accepted 11 August 2019

as comments and sentiments might be useful for firms. Particularly, brand management can be one area of interest, as online reviews might have an influence on brand image and brand positioning.

As pointed out by Liu (2012), firms are increasingly capturing more data about their customer sentiments because opinions are central to almost all human activities and are key influencers of our behaviors. However, regarding the nature of UGC, most extant research employs summary numerical values such as rating scores (Netzer, Feldman, Goldenberg, & Fresko, 2012), while a recent stream of research employs Sentiment Analysis (SA) or text mining with the aim of examining the textual content of UGC and categorizing opinions (e.g. Homburg, Ehm, & Artz, 2015; Liu, Burns, & Hou, 2017). The employment of UGC to answer brand-management related questions faces several challenges. Among them, Liu et al. (2017) highlight two: i) Most big data from social media are textual or graphical in nature, so traditional quantitative analysis is not useful; and ii) researchers need to figure out how to identify brand insights from big data quickly and correctly. These challenges demonstrate the need for a framework that can easily transform textual big data into brand insights (Liu et al., 2017).

Within this context, sentiment analysis techniques are a useful way to examine opinionated text, which contains consumers' opinions towards firms, products, brands, or events. Sentiment analysis is a subfield in Natural Language Processing (NLP), which automatically classifies text by valence (Pang & Lee, 2008). Researchers assume that the text of the message (i.e. the online review) explicitly expresses the consumer's opinion on aspects of firms, brands or products (Liu et al., 2017). While certain techniques split the comments into two classes (negative or positive), other incorporate more sentiment classes (Feldman, 2013). The importance of this technique derives from its ability to understand sentiment polarities from huge volumes of texts and it has previously been employed to analyse review opinions (e.g. Ngo-Ye & Sinha, 2014; Pang & Lee, 2008).

Among UGC, firms might be tempted to manipulate reviews. In fact, fake reviews represent an important challenge for platforms and consumers, who might not be able to differentiate between a true and a real comment and a fake one.

There are commercial systems that try to detect fake reviews, but it is very hard to ensure that a review is fake or not. Some systems analyse both reviews and reviewers considering the number of reviews, purchasing patterns, mismatched dates and other tell-tale signs of suspicious review activity (e.g. a reviewer who is new to Amazon, has posted only one review and uses lots of words like “great” and “amazing” (Brodia, 2018)).

Because it is very hard to discover that a review is fake, some enterprises may hire spammers to post fake reviews to promote their product or discredit products of their competitors. The hired spammers are paid based on the fake reviews (Day, Wang, Chen, & Yang, 2017). Also, they may use robots to do bulk submissions, publishing multiple times with the same or similar text.

In this paper, we focus our attention on Amazon online reviews. Amazon is generally recognized as the most important market place in Western countries (Jindal & Liu, 2007). Every day, thousands of people rely on this online market place to buy products and many of them write an online review about the purchased product. Taking into account that the risk assumed by consumers might be higher when buying online compared to in-store shopping (consumers are not able to see and test the product online), consumers rely on UGC by other consumers as a source of electronic word-of-mouth to make their own decisions. In this sense, brand image is derived not only by signals sent by firms, but also by online reviews written by consumers.

Considering the aforementioned premise, the aim of this paper is twofold:

i. to develop a framework that allows marketing managers to easily interpret qualitative UGC, as it is transformed into quantitative estimates.
ii. to design a modular architecture that combines tools of sentiment analysis and fake review detection to assist marketing managers and consumers in their decision-making process.

Hence, the main contribution of this paper is that to the best of our knowledge, this is the first time that this type of architecture is used in the marketing field.

The results will present all the contributions —including the practical ones (Section 4.3)— extracted from the experimentation section.

This paper is organized as follows: Section 1 provides and introduction on the uses of sentiment analysis and the value of consumer reviews in ecommerce for branding; Section 2 presents a brief literature review of related work on big data techniques applied to marketing, sentiment analysis and fake review detection; Section 3 presents the proposed framework; Section 4 explains the data collection and framework setup and the experimentation and results; and finally, Section 5 provides the conclusions and ideas for future works.

2. Background

This paper deals with the benefits of applying Big Data techniques to Marketing. Hence, an overview of the literature is provided in the following subsection. Our proposal is based on automatic Sentiment Analysis and Fake Reviews, so the following two subsections —2.1 add 2.2— summarize the previous work on these areas. Subsequently, each subsection deals with both aims of this paper separately. Finally, Subsection 2.3 presents the challenges and opportunities found after reviewing previous work.

This section ends with an overview of the challenges and opportunities extracted from this work, which evidences our contribution to the state-of-the-art.

Several authors highlight potential applications of big data to marketing (Amado et al., 2018). In this sense, marketing analytics aims to transform big data from different sources (social media, transactions, survey, sensor network, etc.) into valuable information to sup- port decision-making. Although traditional marketing analytics focuses on improving key performance indicators for better insights regarding advertising, pricing, customer relation- ship management or new product development (Sathi, 2014), nowadays many firms and organizations use big data analytics to follow the stream of information and analyse huge volumes of data in real time. The challenge is to transform the data into insights that managers can use to solve problems in the marketing field and to answer questions such as: what the most interesting product for the market is, how to promote the product for that segment, what communication channel should we employ in that market or what price should we set this week, among others (Amado et al., 2018).

Managing data and extracting from it appropriate information for supporting better decision-making is one of the main challenges for marketers. Although marketers might be used to dealing with data gathered from traditional marketing research techniques (e.g. questionnaires), what is different today is the vast amount of data generated and stored, resulting in the so-called big data revolution (Erevelles, Fukawa, & Swayne, 2016). In fact, big data is currently globally spread and widely accepted. Within this context, the goal of marketing analytics should be the collection, management, and analysis of data to obtain insights into marketing performance, maximizing the effectiveness of instruments of marketing control, and optimizing firms' return on investment (Wedel & Kannan, 2016). Furthermore, potential applications of big data to value creation on classical marketing variables (product, price, place or promotion) have been widely recognized by academics and practitioners (Erevelles et al., 2016). Besides, consumer analytics is one of the most important factors in the big data revolution (Erevelles et al., 2016) as these data provide behavioural insights about consumers. In the same line, Fan, Lau, and Zhao (2015) identify potential applications of big data that lay the foundation for marketing intelligence. For example, in Miah, Vu, Gammack, and McGrath (2017), an application to assist destination management organizations is presented, which analyses the geotagged photos uploaded by tourists in Flickr to predict tourist behavioural patterns at different destinations.

Many websites and market places support eWOM communications providing easy-to-access tools to help consumers offering an online review based on their previous experience. These mechanisms vary from aggregate customer ratings (e.g. numerical (1–5) star ratings) to boxes where consumers can write a text description of their experience. These ratings and comments summarize the individual consumers' evaluations and act as indicators of product quality (Noone & McGuire, 2014; Tsang & Prendergast, 2009). Furthermore, and even more important, they act as a cue to help future consumers to determine product or brand attributes (Sun, Youn, Wu, & Kuntaraporn, 2006). The proliferation of these consumer reviews on the Internet represent a big challenge for producers, as they face the complicated task of analysing this information to provide useful consumer insights that could drive decision-making.

2.1. Sentiment analysis

Sentiment Analysis is an area of study within Natural Language Processing that is concerned with identifying the mood or opinion of subjective elements within a text (Bhadane, Dalal, & Doshi, 2015). It is a growing area given the necessity to understand the people's opinion. With the evolution of Internet and its applications, the textual data increases from many sources. The users publish content and provide vast information from social networks, product review web sites, blogs, and internet forums. The exponential growth of Internet usage has created a new platform where people can freely communicate and exchange ideas and opinions (Rahman & Khamparia, 2016). Specifically, Sentiment Analysis in product re- views is the process of exploring these reviews to determine the overall opinion or feeling about a product (Haddi, Liu, & Shi, 2013).

All the problems that must be resolved in NLP are present in Sentiment Analysis. The work by Jandail (2014) shows six kinds of issues in Sentiment Analysis: 1) In particular domains, a word or sentence can have an opposite meaning. 2) An interrogative sentence or conditional sentence may not have positive or negative sentiment, but a particular word may be. 3) The sarcastic sentences may have the opposite sentiment. 4) Some sentences may have sentiment information, but they do not use sentiment words. 5) An only word can change the feeling polarization in two similar sentences, as well as the fact that for different person, a sentence may have a different sentiment. 6) Natural language Issues Change Place to Place. Regarding the Text Mining approaches required, the work in Demoulin and Coussement (2018) highlights: web mining; classification; clustering; concept extraction; information extraction; and, information retrieval.

Firstly, sentiment analysis classifies product reviews as positive or negative; polarity classification is the basic task. One such application is to use opinion mining to determine areas of a product that need to be improved by summarizing product reviews to see what parts of the product are generally considered good or bad by users (Jandail, 2014). The general opinion about a topic is useful, but it is also important to detect sentiment about individual aspects of the topic (Yi, Nasukawa, Bunescu, & Niblack, 2003). In addition, classifying people based on your opinions or improving recommender systems using the positive and negative customer feedback. Rahman and Khamparia (2016) cited other application domains: shopping to compare products with all descriptions and feedbacks of customers, entertainment to view feedback for movies, business for marketing, research and development, decision-making or political analysis on public feedback.

Cambria, Das, Bandyopadhyay, and Feraco (2016) classified the main existing approaches in four categories: Keyword spotting, lexical affinity, statistical methods and concept-based approaches. The keyword- spotting approach classifies text by affect categories based on the presence of unambiguous affect words. It is popular for accessibility and economy, but it is weak recognizing affect- negated words and it relies on surface features. The lexical affinity approach detects obvious affect words and assigns arbitrary words a probable “affinity” to particular emotions.

Machine learning techniques and statistical analysis had been used, but there has been little use of the fuzzy classifiers in this field especially considering the ambiguity of language and the suitability of fuzzy approaches to deal with this ambiguity (Jefferson, Liu, & Cocea, 2017). They propose a fuzzy rule-based system which can offer more refined outputs using fuzzy membership degrees.

All these techniques need to use a sentiment lexicon. SentiWordNet is a lexicon based in WordNet. SentiWordNet provides each synonym set (synset) of WordNet with three sentiment labels regarding positivity, objectivity and negativity (Hung & Lin, 2013). Other word lists are Affective Lexicon, WordNet-Affect, SenticNet and Afinn (Nielsen, 2011). Afinn is interesting because it is an affective word list manually rated for valence with an integer between −5 (negative) and +5 (positive), providing emotional ratings for 2476 English words. Many researches are focused in analysing product reviews to get feedback about the product and make decisions. García-Moya, Anaya-Sánchez, and Berlanga-Llavori (2013) propose a new methodology for the retrieval of product features and opinions from a collection of free-text customer reviews about a product or service. Also, (Singla, Randhawa, & Jain, 2017) classifies the text in positive, negative and includes sentiments of anger, anticipation, disgust, fear, joy, sadness, surprise and trust. Products were rating based on sentiment analysis, processing automatically textual reviews and classifying them according their polarity confidence (Sindhu, Vyas, & Pradyoth, 2017). Paknejad (2018) studies different machine learning approaches to determine the better options for sentiment classification problem for online reviews using product reviews from Amazon.

Many sources of datasets have been used for this work. For example, He and McAuley (2016) use Amazon's product reviews. Also, IMDB review for movies can be used to analyse sentiment, featuring 25,000 movie reviews for training and testing (Maas et al., 2011). Sentiment140 is a dataset with 160,000 tweets (Go, Bhayani, & Huang, 2009).

Also, we found previous work analysing data on social media, for example He, Wu, Yan, Akula, and Shen (2015) use sentiment analysis to identify the leading companies in the technology or retail sector, in relation to social media comments. The objective being to highlight the areas where a company is perceived to be excelling or showing a potential problem area that needs to be addressed. In the case of Twitter, the use of slang and emoticons, limitation of 140 characters by tweet and misspellings force to study the pre-processing of text of tweets (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011). We found a bootstrap ensemble framework for Twitter sentiment analysis to build sentiment time series that are better able to reflect events eliciting strong positive and negative sentiments from users. The use of machine learning for Twitter sentiment analysis is wide (Hasan, Moin, Karim, & Shamshirband, 2018; Jain & Dandannavar, 2016; Neethu & Rajasree, 2013).

There are several tools to work with sentiment analysis: IBM (www.ibm.com/analytics)., SAS (www.sas.com/social), Oracle (www.oracle.com/social), SenticNet (www.business.sentic.net) and Luminoso (www.luminoso.com). Most tools are limited to a polarity evaluation or a mood classification and they cannot capture opinions and sentiments that are expressed implicitly (Cambria et al., 2016). Language R has specialized libraries to work with text data and combining it with review datasets is a simple and powerful tool for sentiment analysis. Use of R in sentiment analysis is wide (Liske, 2018; Paracchini, 2016).

2.2. Fake reviews

Internet technologies have totally changed the way people buy products. The behaviour of the consumer has also changed, with consumers writing online reviews about the products they acquired —UGC—. As stated above, when a customer wants to buy a product online they take time to check the product scores and read opinions. They are less likely to choose a product based solely on price as they may prefer the product with a higher score and excellent reviews. A product with more positive reviews has more chance than products

with negative reviews. Unfortunately, the importance of the review is misused by certain parties who tried to create fake reviews, both aimed at raising the popularity or to discredit the product (Wahyuni & Djunaidy, 2016). This practice is called review spam (Heydari, Tavakoli, Salim, & Heydari, 2015).

In near future, spam reviews might damage the entire online review systems and finally could cause a gradual loss of credibility. Hence, the first step towards securing the online review system is detecting the spam reviews. (Mahalakshmi, 2017).

Reviews on brand and non-reviews are relatively easy to detect manually, so they can be used on traditional classification learning. However, for untruthful opinions, manual labelling by simply reading the reviews is very hard (Jindal & Liu, 2007). These authors also analysed what kinds of reviews are harmful and are likely to be spammed. Fake positive reviews for good quality products are not harmful, but fake positive reviews for bad quality products or fake negative reviews for good quality products are harmful.

Also they found a large number of duplicate and near-duplicate reviews. The following types of duplicates including near-duplicates: 1. Duplicates from different userids on the same product. 2. Duplicates from the same userid on different products. 3. Duplicates from different userids on different products. In this research to detect duplicate reviews, they used 2 g based review content comparison. The similarity score of two reviews is the ratio of intersection of their 2 g to the union of their 2 g of the two reviews, usually called the Jaccard distance. Review pairs with similarity score of at least 0.9 were chosen as duplicates.

To build a model, they created the training data defining a large set of features to characterize reviews. They propose three types of features: review centric features, reviewer centric features, and product centric features. Some review centric features are: Number of feedbacks, number of helpful feedbacks and percent of helpful feedbacks, length of the review title and length of review body. They found a large number of duplicate and near-duplicate reviews written by the same reviewers on different products or by different reviewers (possibly different userids of the same persons) on the same products or different products.

Another challenging task is to identify a list of spamming signs and indicators. Strategies are designed with the facts extracted from the consumer studies and explicit domain knowledge for distinguishing genuine and spamming reviews. There are few assumptions in the literature and researches: Genuine author will not write multiple review texts with ratings and smaller reviews are more genuine than larger texts (Mahalakshmi, 2017).

Traditionally, review spam detection problems can be formulated as machine learning tasks (Jiang, Cui, & Faloutsos, 2016; Shivagangadhar, Sagar, Sathyan, & Vanipriya, 2015; Zhang, Wu, & Cao, 2018). Three main categories of detection methods have been used: supervised; clustering; and, graph-based methods. Supervised methods infer a function from labelled information (reviews in our study). Regarding supervised approaches, Li, Huang, Yang, and Zhu (2011) first describe the influence of different features in the supervised learning framework using a manually built spam collection. They observe that the review spammer consistently writes spam and they can identify if the author of the review is spammer.

With meta-data about the review, Farooq and Khanday (2016) mine many types of abnormal behavioural patterns of reviewers and their reviews. For example, reviewer wrote only positive reviews for a brand and negative reviews for a competing brand. Also, they can analyse product description or sales volume/rank. A product with low sales but many positive reviews is hard to believe. Furthermore, the authors analyse the main features used in supervised learning approaches.

Holla and Kavitha (2018) show some of the typical characteristics of fake reviews: Less information about the reviewer, review content similarity, short reviews, sudden uploading of reviews in the same time frame, focus on personal information, and excessive use of positive and negative words. This paper discusses the various techniques used for identifying fake reviews: Detection of Fake Review created by Groups, generation of synthetic reviews and their detection, detecting spam review through sentiment analysis, neural network used to detect fake reviews by exploiting product related review features, fraud detection in online reviews by network effects, detecting fake reviews by the principle of collective positive unlabelled learning, word order preserving convolutional neural network used for spam detection, and detecting singleton spam reviews.

Elmurngi and Gherbi (2017) analyse online movie reviews using sentiment analysis methods and machine learning algorithms in order to detect fake reviews. They compare five supervised machine learning algorithms: Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN-IBK), KStar (K*) and Decision Tree (DT-J48) for sentiment classification of reviews using a dataset of movie reviews. The measured results show that the SVM algorithm outperforms other algorithms, and that it reaches the highest in detecting fake reviews. Other approaches which used SVM with excellent results were presented in (Mukherjee, Venkataraman, Liu, & Glance, 2013; Ott, Choi, Cardie, & Hancock, 2011).

Tavakoli, Heydari, Ismail, and Salim (2016) proposed an effective framework to be used for spam detection research. They discussed the categorization and fundamental explanations of some of the elements and factors involved with spam detection. Finally, they collected the features that were used in proposed review-spam detection techniques. These features were classified in review-based detection metrics, product-based detection metrics and spammer detection.

A principled hybrid learning model called hPSD to combine both user features and user- product relations for spammer detection is proposed by Wu, Wang, Wu, Cao, and Zhang (2015). hPSD, includes feature discretization, reliable negative set extraction and hybrid learning scheme. Extensive experiments are conducted on both movie data with shilling injection and Amazon data with true yet hidden promoters, to validate the effectiveness and practical value of the proposed model.

A significant number of supervised proposals are based on linear/logistic regression models (Jindal & Liu, 2007; Jindal & Liu, 2008; Lim, Nguyen, Jindal, Liu, & Lauw, 2010). In general, linear regression models the relationship between a scalar dependent variable (label) y and one or more independent variables (features) x. Logistic regression assumes a logistic function to measure the relationship between labels and features. Other approaches are rule-based or use decision trees to detect review spam (Gao, Xu, & Wang, 2015; Jotheeswaran & Kumaraswamy, 2013; Saumya & Singh, 2018). Finally, proposals using naive Bayesian models, which assume strong independence between features, have been defined (Li, Ott, Cardie, & Hovy, 2014).

The second category within the machine learning class of spam detection methods is based on clustering. A brief summary could be that clustering is the task of grouping a set of objects so that those in the same cluster are more similar to each other than to those in other clusters. These works (Ha, Vu, Pham, & Luu, 2011; Heydari, Tavakoli, & Salim, 2016; Jia, Zhang, Xia, Zhang, & Yu, 2010; Liu & Pang, 2018; Mukherjee et al., 2013; Mukherjee, Liu, & Glance, 2012) have obtained good results.

Finally, the graph-based methods (third category) use graphs to represent interdependencies through the links or edges between objects via network information in social spam and behavioural information in link-farming scenarios. Graph-based detection methods can be categorized into PageRank-like approaches and density-based methods. PageRank-like approaches solve a suspicious node detection problem in large graphs from the ranking perspective (Akoglu, Chandy, & Faloutsos, 2013). Density-based detection methods in graphs look for areas of higher density than the remainder of the graphs/data (Ye & Akoglu, 2015).

A summary of the characteristics and performance of the main methods described in this section will be presented in Section 4.2.5, where our proposal of a fake-review detection tool is evaluated and compared with previous works.

2.3. Challenges and opportunities

After reviewing the previous work, some findings can be drawn:

- Every day new reviews are generated for online products. Stores like Amazon have a textual big data that contains valuable information for marketers which is not being used. It is important to extract from those opinions, the signals that can facilitate better decision-making along with quantifiable variables such as price or star rating.
- NLP tools such as sentiment analysis help quantify a reviewer's opinion. Many of these works are used to analyse opinions in social networks and online reviews. The score obtained will represent the evaluation of a variety of product characteristics and may be slightly different than the general opinion of the product. It is important to understand the relationship between the star rating and the opinion, as this is can be used to evaluate various characteristics of the product or service.
- In online reviews, fake reviews may affect the classification of a product. For example, in a website for tourism, there are fake reviews that can provoke loss of prestige for the company (hotel, restaurant, etc). If we eliminate the fake reviews, the actual rating will adjust and will offer a more accurate (higher) estimate. Another example is that fake reviews alter the results of technological product ratings. If we eliminate the fake reviews, users can decide their purchases in a more reliable way, but the problem is identifying with certainty whether a review is false. There are some proposals to discover cases of fake reviews, but they have limited accuracy. However, using NLP technology we can discover certain relationships between opinions by their similarity that lead us to suspect that they are not reliable. In this case, it would be better to withdraw them or reduce their opinion weighting.

3. The proposed framework

In this section we present our proposal, the Fake Review Detection Framework—FRDF—(Fig. 1). The FRDF is based on the use of Sentiment Analysis and fake review detection tools. Essentially, additional information is extracted from user reviews to modify, if necessary, the original star score assigned by users. FRDF can inform and thereby improve decision-making of marketing managers, resellers and consumers.

As shown in Fig. 1, 4 stages are described: (1) review pre-processing, (2) Sentiment Analysis, (3) Fake Review Detection, and (4) dashboards.

This framework receives as input big data product reviews and the relevant information of each product, such as: price; the brand; and, the categories of the product. This data is analysed to extract new market intelligence that will help managers and consumers to make more accurate decisions regarding products.

User reviews usually contain a score and a comment in unstructured text. The rating is a score that the reviewer chooses in a range, for example, from 1 to 5 stars, where a value of 1 is a very bad rating and a value of 5 is a very good rating. However, some discrepancies can occur when, for example the product is given an overall score of 4 or 5 stars, but the user review indicates criticism of specific product characteristics. In this case, the FRDF will adjust the rating accordingly, even if the textual comment has both positive and negative opinions of different aspects of the product.

The FRDF aims to use different tools of Sentiment Analysis and Fake Review Detectors. This framework must be sufficiently modular to allow the incorporation of diverse sentiment analysis and fake review detection tools. These may include a set of NLP techniques —specifically lexical analysis, syntactic analysis, semantic analysis, etc.— The sentiment analysis tools enable a quantitative ranking to be obtained from the textual reviews. The sentiment score can be used as an additional criterion to search for the best products within a product category or within a brand of products.

The reviews that are considered possible fake reviews will be ignored in our framework. We use NLP techniques to detect similarity between reviews. Although it cannot be guaranteed that a review is false, it is more likely that two very similar reviews that come from different reviewers are false. Each review is compared to all other reviews and the similarity between them is calculated. If the similarity exceeds a certain threshold, they will be labelled as fake reviews. The fake review detection provided by this framework allows a comparative analysis of the product ranking that includes and excludes the fake reviews, thereby providing real-time and more accurate market intelligence for decision-making.

The four stages visualized in Fig. 1 are described in detail next as well as the theoretical background of the methods applied.

1) Review Pre-processing involves dividing the sentences of each user review into tokens and tagging them with lexical, syntactic and semantic information. These tags or labels will improve the user review processing results. General Stop_Words and specific e-commerce terms are removed.
2) Sentiment Analysis seeks to find, based on the words used, the degree to which the comment about the product is positive or negative. It is highly likely that a product with a high star rating will also have mostly positive comments; and the reverse will apply with a product that has a low star rating.
   Among the most important set of NLP tools for research purposes are: Freeling; Stanford CoreNLP; OpeNER; Tidytext; and, other libraries for text analysis using R language.

In our research we have used R (Team, 2013) and some libraries for text analysis, mainly Tidytext (Silge & Robinson, 2016). The R language is a programming language widely used in data mining that allows

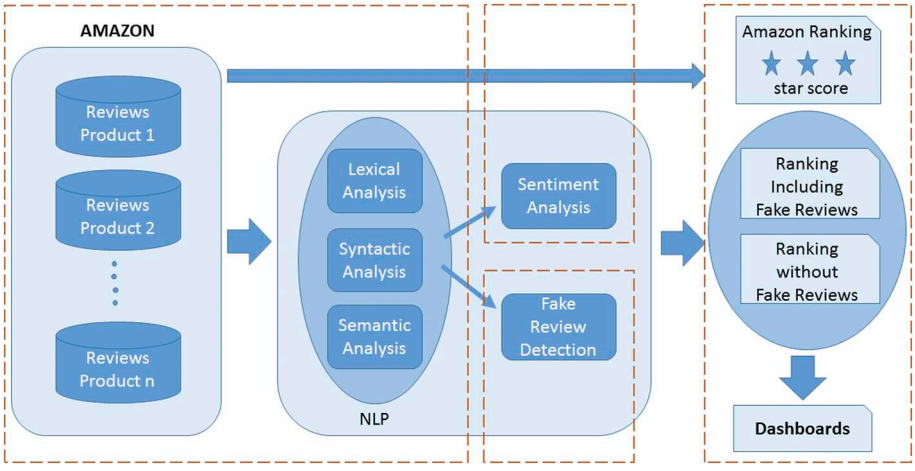

Fig. 1. General framework for enriching product information using NLP, sentiment analysis and fake review detection tools.

statistical analysis. It has different libraries with calculation functions and graphics. Particularly, it has a set of libraries that facilitate the processing of natural language.

Tidytext is a package developed to work on text that follows the principles of tidy data, making text processing easier and consistent with tools that are already used (Silge & Robinson, 2016).

The Afinn Sentiment Lexicon was used for sentiment analysis in our experiments. This lexicon is a list of 2476 words that express some degree of sentiment was used. This list of words, manually constructed by Finn Arup Nielsen between 2009 and 2011, contains a qualification of how positive or negative a word is, using a range from −5 (negative) to +5 (positive).

Any sentiment analysis tool that returns a sentiment qualification in a polarity range can be used in this Framework. This range can be normalized to an expected range, for example [1–5]. Particularly in this research, we used the following lexical affinity approach that was used by other researches (Liske, 2018; Paracchini, 2016):

Input: Sentiment lexicon list, Stop_Word list.

1) The review-text is split into individual words (Review_Words).
2) The words that belong to Stop_Word list are removed from Review_Words.
3) Each word in Review_Words is searched into the Sentiment lexicon list. If the word is in the list, add the sentiment rating of this word.
4) The sentiment value of the user's review is the mean of all these sentiment ratings.

Eq. (1) shows the overall mathematical equation to obtain the sentiment value of a review.

$$sentiment(Review) = \frac{\sum_{w \in R} emotional\_rating(w)}{|R|} \quad (1)$$

where $R = Review\_Words \cap Afinn\_Words$

5) The value range of this expression can be normalized to the expected range.
6) This value is added to the review information.

This model is based on a polarity sentiment word list. This approach is simple and fast to apply, but it has some of the problems mentioned by Jandail (2014). We implement this algorithm using language R and Tidytext library.

The framework allows for the use of other sentiment analysis tools and then calculates the average sentiment value.

The other sentiment analysis tool used for this research was Stanford CoreNLP, an extensible pipeline that provides core natural language analysis (statistical NLP, deep learning NLP, and rule-based NLP tools). It provides sentiment analysis with a compositional model over decision trees using deep learning. Nodes of a binary decision tree of each sentence, including, in particular, the root node of each sentence, are given a sentiment score (Manning et al., 2014). CoreNLP is a classifier whose results reflect the distribution of the corpus. With training, it is useful for building a domain-specific model. It detects 5 possible classes of sentiment classification: very negative, negative, neutral, positive, and very positive.

3) A fake review can alter the statistics and opinions of a product, sometimes favouring it and sometimes damaging it. A fake review can be done by a person who is hired to write comments about a product. Robots can also be used to automatically insert comments about products.

For this reason, it is very important to detect and remove fake reviews in order to obtain more reliable results. The problem is to know when a review is fake, and there is no total certainty that a review can be fake. As previously described in the literature section, there are many techniques to detect fake reviews. In our case study, we will use an approach to calculate the *cosine similarity measure* (Manning, Raghavan, & Schutze, 2008). Once the similarities between all the reviews have been calculated, the reviews that exceeded a specified threshold are removed.

Our Framework can use different fake review detection techniques. Particularly, FRDF uses a strategy similar to Lau et al. (2011), where a review is considered fake when we find two or more very similar reviews from different reviewers or for different products.

The *similarity measure* will be a numerical value of the similarity between two elements. This measure is in the range from 0 to 1 (0: lower, 1: higher). We have defined a threshold to determine when a review is considered fake. If the similarity measure between two reviews is above that threshold, it is a candidate to be considered a fake review and it would be removed from the valid reviews.

The similarity measure helps to find out similar product comments. The cosine similarity function is an effective way to compare two reviews. We have implemented the tf-idf-cosine similarity which is shown in the Eq. (2).

$$S(Q,D) = \frac{\sum_w Q_W D_W}{\sqrt{\sum_w Q_w^2}\sqrt{\sum_w D_w^2}} \quad (2)$$

where

$$Q_w = \frac{tf_{w,Q}}{tf_{w,Q} + \frac{k\ |Q|}{avg\ |Q|}} \cdot \log \frac{|C|}{df_w}$$

$$D_w = \frac{tf_{w,D}}{tf_{w,D} + \frac{k\ |D|}{avg\ |D|}} \cdot \log \frac{|C|}{df_w}$$

With this calculation, we compare the reviews to determine the closeness between each pair of them. If a similarity threshold is exceeded between two reviews, it could be considered that the reviews are the same. The reason could be either because the reviews were made by a bot or because there is someone who wanted to improve or worsen the general opinion about a specific product.

The process for analysing the similarity of reviews is as follows:

1. We remove commonly used words which are not relevant in the comparisons (Stop_Word).
2. We also remove commonly used words in online sales product reviews (e.g. buy, bought, price, product, etc.).
3. We obtain word roots by using Porter's Stemmer algorithm, in such a way that variations of the same word are considered the same. For example, a word in both plural and singular form is considered the same, or different conjugations of a verb are considered the same word.
4. We calculate several measures related to term frequencies: frequency of each term within each review; number of times each term appears throughout the corpus; number of reviews in which each term appears; and number of comments throughout the corpus.
5. We obtain the similarity value between two reviews using Eq. (2).
6. If the similarity value is greater than the specified threshold, we must check if the two reviews are from two different users or about two different products. If neither of these cases applies, then both reviews are removed. Only the remaining reviews will be considered for marketing analysis.

After finishing stage 2 (sentiment analysis) and stage 3 (fake review detection) we will obtain a new score with the sentiment analysis and the data refined with the fake review detection. Subsequently, we will carry out a series of analyses that can help the decision-making process for sellers, manufacturers and users.

In our framework, we work with a series of questions that could be answered with the extensive information that is handled in big data of product reviews. We will consider the star score, the score obtained with the sentiment analysis and the product price. Examples of these questions are the following:

Table 1
Top ten product classification by categories.

#	Category	Amount of products
1	Computers & accessories	1445
2	Camera & photo	1280
3	Accessories & supplies	773
4	Cables & accessories	687
5	Audio & video accessories	588
6	Accessories	472
7	Cables & interconnects	355
8	Portable Audio & Video	263
9	Digital cameras	219
10	Point & shoot digital cameras	183

- What is the best product in each category according to price or consumer satisfaction?
- What is the product with the highest satisfaction of a specific brand?
- Considering that customer satisfaction is important, a seller might ask: If my product is the most satisfactory according to customer opinion, what would be the best price to generate a sale?
- For a customer who is looking for the best product at the best price within a category, what is the best product in terms of price and customer satisfaction, not only according to the star score, but also according to customer opinion?

4) In the last stage of FRDF, we present the dashboards that will help in the decision-making process. Using these dashboards managers can view products that require decisive action. For example, a product may have limited sales due to negative sentiment towards the product or, alternatively, the product may a positive sentiment value, but is priced too high to generate new sales.

4. Case study

In this section, a case study of Amazon, the commerce social network, is presented by focusing on a set of product reviews sold on the Amazon website. Next, in Subsection 4.1, the dataset is described, and then, in Subsection 4.2 all the experiments carried out are explained. To conclude this section, Subsection 4.3 summarizes the theoretical and practical contributions of this work following the experimentation.

4.1. Data collection and framework setup

The data corpus is obtained from “Amazon Product Data by Julian McAuley” in http://jmcauley.ucsd.edu/data/amazon/, visited on 14th of January 2019 (He & McAuley, 2016).

This dataset contains product reviews and metadata from Amazon from May 1996 to July 2014. Each review includes a score from 1 to 5 (star rating – 1: lower, 5: higher) and a comment given by the user. Also, for each product, there is a dataset with product-specific information, namely, product description, category information, price, brand, and image characteristics. All the reviews refer to 4181 different products. These products can be classified into 490 different categories. Table 1 shows the top ten categories.

The dataset of these product reviews has the following structure shown with a sample:

{
"*reviewerID*": "A2SUAM1J3GNN3B",
"*asin*": "0000013714",
"*reviewerName*": "J. McDonald",
"*helpful* ": [2, 3],
"*reviewText* ": "I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!",
"*overall* ": 5.0,
"*summary*": "Heavenly Highway Hymns",
"*unixReviewTime*": 1252800000,
"*reviewTime*": "09 13, 2009"
}

As can be shown, for each review, we can distinguish a comment (*reviewText*) and a score of the consumer about the product (*overall*). Furthermore, the following metadata for all the products is available:

Data analysis was done using the R language. Following the

{
"*asin*": "0000031852",
"*title*": "Girls Ballet Tutu Zebra Hot Pink",
"*price*": 3.17,
"*imUrl* ": "http://ecx.images-amazon.com/images/I/51fAmVkTbyL._SY300_.jpg",
"related":
{
"*also_bought* ": ["B00JHONN1S", ..., "B00BFXLZ8M"],
"*bought_together* ": ["B002BZX8Z6"]
},
"*salesRank* ": {"Toys & Games": 211836},
"*brand* ": "Coxlures",
"*categories*": [["Sports & Outdoors", "Other Sports", "Dance"]]
}

abovementioned model, the sentiment word list used was Affin Word List. The used Stop_Word are listed in Tidytext library for R. We built the specific commerce domain word list based on the most repeated words on this Amazon review corpus. With these inputs, the evaluation of each review using the Eq. (1), will fix the value range of this expression from −5 to +5. This value is then added to the review information.

In order to understand the sentiment analysis calculation, the following review is presented:

“We got this GPS for my husband who is an (OTR) over the road trucker. Very Impressed with the shipping time, it arrived a few days earlier than expected... within a week of use however it started freezing up... could of just been a glitch in that unit. Worked great when it worked! Will work great for the normal person as well but does have the” trucker" option. (the big truck routes - tells you when a scale is coming up ect…) Love the bigger screen, the ease of use, the ease of putting addresses into memory. Nothing really bad to say about the unit with the exception of it freezing which is probably one in a million and that's just my luck. I contacted the seller and within minutes of my email I received a email back with instructions for an exchange! VERY impressed all the way around!”

Using R language and Afinn, a sentiment value of 2 was obtained. This score can be classified as positive. The original reviewer assigned a score of 5 stars to this product.

This example shows that, although the overall rating on the product's star score is very positive, there are specific details that were criticized (“…Nothing really bad to say about the unit with the exception of it freezing which is probably one in a million and that's just my luck…”). For that reason our sentiment analysis module obtains a score lower than star score.

To normalize the values of sentiment analysis and star rating by the user, the following mapping is done:

If sentiment value is >= 3 then it is very positive and has 5 stars
else if sentiment value is >= 1 then it is positive and has 4 stars
else if sentiment value is >= -0.5 then it is neutral and has 3 stars
else if sentiment value is >= -3 then it is poor and has 2 stars
else the sentiment is very poor and has 1 star.

As mentioned above, our framework is modular and able to work with different sentiment analysis tools. For example, if we use Stanford CoreNLP with the previous review, we calculated an average of the sentiment values of each sentence and we obtained a neutral sentiment value (equivalent to 2.5 stars).

After obtaining the sentiment score, the following step is to analyse the correlation between sentiment and star scores. The consumer can mention many ideas in the textual review that cannot be reflected in the numerical one. The numerical rating (stars) is global for the product or the experience of using it, but there are always some characteristics of the product that the user does not like. This can cause a difference between the star score and the sentimental value, which would be reflected in a lower correlation between the two scores.

Two methods of correlation analysis were used. Pearson's correlation coefficient was applied to measure the degree of variation between the star score and the sentimental value. However, since we have different value ranges between the star score [1–5] and the sentimental score [−5–5], we also used a rank correlation, specifically the Spearman correlation.

4.2. Experimentation

According to the challenges and opportunities presented in Section 2, we describe in this section the experiments carried out. First of all, we introduce the correlation analysis between star and sentiment scores performed. After that, the experiments with sentiment analysis and fake review detection tools are described. Furthermore, the performance of the used tools are detailed.

4.2.1. Correlation between star and sentiment values

In this experiment we accomplished the correlation analysis between the star and sentiment scores. Firstly, we calculated the mean of the star scores by product. In Fig. 2, we show the review distribution based on the star score and the average star score by product. As can be seen, a high percentage of evaluations were rated with the highest score. Most of the products were well qualified, and 184 products had perfect qualification.

After that, the sentiment analysis of our framework was applied on the 100,000 selected reviews to obtain their sentiment values. Some reviews were previously removed as they did not have any words with sentiment rating. The remaining reviews amounted to 95,737. By making a statistical analysis of the sentiment score of each review, most of the reviews obtained a score of 4 (see Fig. 3). The same analysis was done for the products based on the sentiment score. Comparing the charts between the reviews with star score and sentiment value, some differences can be observed. Most reviews received a star score of 5, whereas most reviews received a sentiment value of 4. This indicates that, although in general terms they are very good products or the experience with the product was excellent, there were specific details that were criticized and, therefore, some negative comments were mentioned. For this reason, we decided to analyse these results in greater detail by calculating the correlation between both variables. We obtained the Pearson correlation coefficient with a value of 0.31, whereas the Spearman correlation was 0.28. These low correlation coefficients confirmed that there was some information in the textual sentiment analysis that was not present in the star score. Including this information is crucial for having better comparison criteria between products.

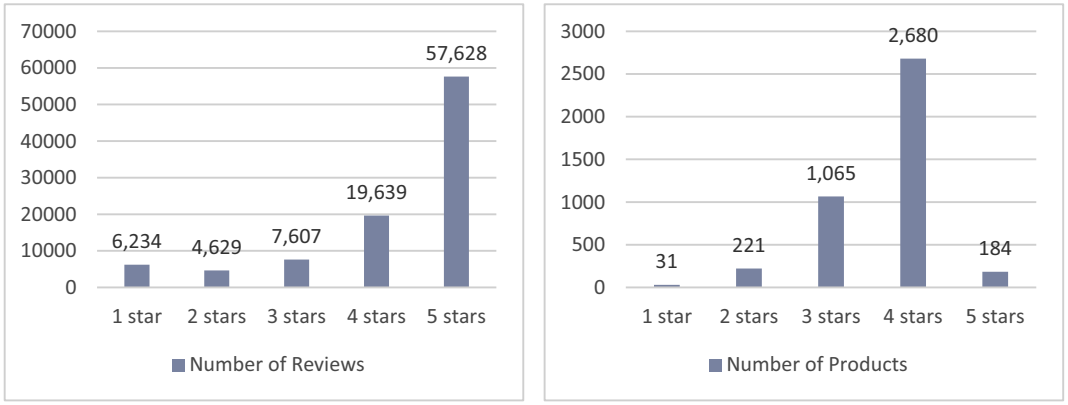

Fig. 2. Classification of reviews and products by star score.

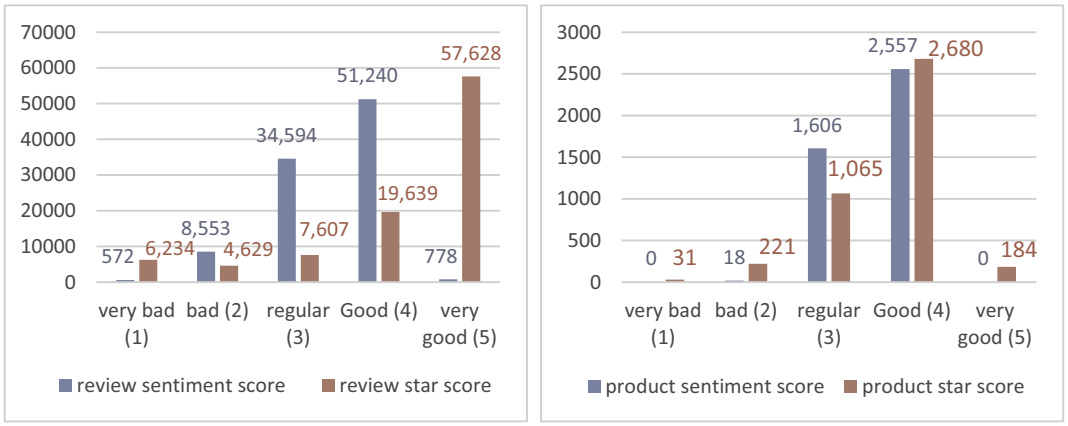

Fig. 3. Classification of reviews and products by star and sentiment score.

Table 2
The top three products in the Headsets and Microphones category based on an average score.

Product	Price	Star score	Sentiment value	Average score
B00004T8R2 (Panasonic)	7.38	4.38	1.45	0.8172
9862510447 (Generic)	5.00	3.66	1.78	0.7910
B00005ML7Q (Koss)	11.95	4.15	1.44	0.7658

We can corroborate our conclusions with the following three examples extracted from the data corpus. The first review was scored with 5 stars, but it was scored with 3 points using sentiment analysis. This fact can be justified due to the global opinion about the product, which was very positive (a high star score), but some minor negative characteristics of the product were expressed (for instance, the user commented "The downside is occasionally the USB hotsync doesn't work right, requiring you to try a sync again"). The user review is the following:

> *"For anyone who's wished for an organizer that can do it all for them. Get this one! For something merely pocket sized you get a memopad, to-do list, datebook, contact book, and the ability to do so much more! I find myself using this thing to remember key appointments in my life. Also, when I'm bored somewhere, I fire up one of the games I've downloaded, or using AvantGo (downloaded program) to check on the news lately. The downside is occasionally the USB hotsync doesn't work right, requiring you to try a sync again. Once in a great while having to reboot your computer to get it to work. However, this happens few and far between (maybe 1 in 20 hotsyncs). However, hotsyncs are so easy, just push a button and voila you're synced up with Outlook! In short, anybody who's got a life they consider complex should consider getting one of these. You won't regret it!"*

The second example was scored with 5 stars and 2 sentiment points. In this review the user wrote "While the Visor is a stable product, it is possible to completely confuse it and end up losing all of your data". Next, the review is shown:

> *"While the Visor is a stable product, it is possible to completely confuse it and end up losing all of your data. I did it while trying to wirelessly synchronize with Omnisky and I ended up losing all of my data by doing a hard reser (starts your Visor over from scratch). The first time, it was quite a blow because I had to reinstall everything. The next day I ordered this product and when it arrived I immediately made a backup. I am thinking about getting two of these items, one for a baseline backup of my system and one for my everyday backup of essential information. It is quick and easy, when you insert it into the slot the Visor loads the program and prompts you for action. You simply click on the icon that identifies backup or restore and it does the rest. When it is done you can remove the module and everything is back to normal."*

However, it is important to emphasize that there are other problems related to the Lexical Affinity Sentiment Analysis methodologies (Cambria, Schuller, Xia, & Havasi, 2013; Jandail, 2014) which influence the low correlation between the two scores. For example, we can mention two problems: the negation phrases and the ambiguity in terms that could refer to positive or negative sentiments. The third example shows a review which contains negations: "it wasn't too hard to install" or "Nothing bad to say about it". The third review is the following:

> *"It does the job and it wasn't too hard to install so i'm a happy tv viewer. Nothing bad to say about it. Came with a bunch of bolt options for different make and model TV's."*

Table 3
The top three products in the eBook Reader category based on an average score.

Product	Price	Star score	Sentiment value	Average score
1400501776 (Barnes & noble)	89.99	4.16	1.22	0.6877
140053271X (Barnes & noble)	79.49	3.83	1.15	0.5778
1400532655 (Barnes & noble)	113.99	3.81	1.09	0.4158

Table 4
The five best products of the Sony Brand sorted by customer satisfaction.

Product	Categories	Price	Star score	Sentiment score	Average score
B00008AYBH	Accessories & supplies blank media DVD-R discs	2.0	4.75	2.71	3.7292
B00009RUFZ	Accessories cables & cords camera & photo	13.15	4.67	1.89	3.2793
B0001MQUNS	Accessories batteries batteries & charges camera & photo	2.90	4.87	1.68	3.2778
B000067S9H	Accessories & supplies blank media CD-RW discs	17.13	4.60	1.83	3.2167
B0001AU7SY	Accessories batteries batteries & charges camera & photo camera batteries	20.25	4.87	1.49	3.1823

4.2.2. Comparative review mining

One of the main aims of our framework is to assist managers and consumers in the decision-making process related to the reviewed products. In this section, we detail the process for providing additional information to the star score after performing the sentiment analysis.

For example, the decision maker tries to answer the following question: “What is the best product by price or user satisfaction in a specific category?” Our framework easily shows the best products of each category. Any system could also obtain this information from structured data sorted by price or star score. However, our proposal provides additional information sourced from the analysis of the sentiment value.

In our approach we will use three variables: (1) the star score, (2) the sentiment value, and (3) the price. We assigned different weights to the variables in order to obtain the average score and, therefore, specify the best product. This score was calculated from the weighted average between normalized price, star score and sentiment value using Eq. (3). A weight of 0.4 was assigned to the price, and 0.3 to each of the star and sentiment scores. The price is normalized using the maximum and minimum price, star score and sentiment score into the same category.

$$\begin{aligned} & Normalized\,\mathrm{Pr}\,ice\,(product) * 0.4 + AvgScore\,(product) \\ = \; & NormalizedStarScore\,(product) * 0.3 + \\ & NormalizedSentimentScore\,(product) * 0.3 NormalizedPrice\,(product) \\ = \; & \frac{Max\,\mathrm{Pr}\,ice - \mathrm{Pr}\,ice + 1}{Max\,\mathrm{Pr}\,ice} NormalizedStarScore\,(product) \\ = \; & \frac{StarScore - MinStarScore}{MaxStarScore - MinStarScore} NormalizedSentimentScore\,(product) \\ = \; & \frac{SentimentScore - MinSentimentScore + 1}{MaxSentimentScore - MinSentimentScore} \end{aligned} \tag{3}$$

In Table 2, we can see the three best products of the category “Headsets and Microphones”. We should remark that the best product, B00004T8R2 (Panasonic), was not the most expensive one, although it had the best user reviews according to sentiment value.

In Table 3, the three best products in the “eBook Reader" category are shown. Again, the best product was not the most expensive one, according to the average score obtained. However, the product with the best star score and sentiment value was more expensive than others. In this way, the buyer has enough information to choose between the best ranked product according to average score and a product with a lower price, which has an average score that is close to the best performing product.

Table 6
Confusion matrix to evaluate the sentiment analysis tool.

	Real positive	Real negative	Real neutral
Predicted positive	46,753	2304	2961
Predicted negative	5183	2954	988
Predicted neutral	25,331	5605	3658

Table 7
Performance of sentiment analysis tool.

	Accuracy	Recall	F-measure
Positive	0.898785	0.605084	0.723255
Negative	0.323726	0.271932	0.295577
Neutral	0.105741	0.480873	0.173361
Average	0.442751	0.45263	0.397398

We carried out a similar analysis to find out the best product within a category. This information can be important for the decision of a buyer because he can visualize the price and the clients' satisfaction and make the decision based on the three mentioned variables.

The previous analysis can be repeated by looking for the best product of an established brand according to the consumers' review: “What is the most satisfying product for a given brand?”. A company can find its main product not only for sales but also for customer satisfaction, which is evaluated as a previously described average score from the star score and the sentiment value. In Table 4, the five most satisfying products of the Sony brand are presented.

Our analysis also finds the best products according to average score that do not have the best price. Analysing the best price for similar products in the same category helps companies to determine optimum competitive pricing strategies.

Table 5 shows some examples of products that are the best rated and not the most expensive in their category. Our framework extracts these cases so that decisions can be made about in relation to increasing the price, or to produce a product that maintains quality, but can compete

Table 5
Best scored product by categories.

Category	Product	Brand	Mean star score	Mean sentiment score	Mean average score	Price	Best price	Worst price
Camera	B0001AUAUE	Sony	4.75	1.58	0.6840	29.99	16.99	64.95
I/O Port Cards	B0000E2Y7Q	startech	4.80	1.37	0.5457	37.34	9.97	58.99
Backpacks	B00020BJA8	Targus	4.60	1.26	0.2022	52.99	39.95	64.34
Keyboards & styluses	B00004TF4V	landware	4.80	1.56	0.5295	39.99	6.06	39.99

Table 8
Three types of duplicate spam reviews.

Spam review type	Number of review pairs
Different userids on the same product	95
Same userid on different products	855
Different userids on different products	1250

Table 9
Fake reviews vs. product rating.

	Positive fake review	Negative fake review
Good rated product	1220	79
Bad rated product	1	1
Average rated product	12	15

with market leading products due to their lower price, through more efficient and innovative production methods.

4.2.3. Performance of sentiment analysis tool

We have used the star score as the real sentiment value to evaluate the performance of the sentiment analysis. We evaluated manually 1000 reviews to take this decision. (Denecke, 2008) used the star score to classify real positive and negative reviews. Both star and sentiment scores were normalized to Positive, Negative and Neutral values. After this, we built a confusion matrix shown in Table 6 to calculate the performance measures.

The measures were calculated as:

$$\text{Accuracy} = \text{True Positive}/(\text{True Positive} + \text{False Positive})$$

$$\text{Recall} = \text{True Positive}/(\text{True Positive}+\text{False Negative}+\text{False Neutral})$$

$$F1 = 2 * \text{accuracy} * \text{recall}/(\text{accuracy} + \text{recall})$$

The performance is reported in Table 7 with the accuracy, recall, and F-measure for each polarity classes as the overall performance. This choice is due to the standard methodology adopted for benchmarking of sentiment analysis systems in evaluation campaigns (Novielli et al., 2018).

The results achieved 89.87% accuracy for finding the positive sentiment. This means that of 52,018 reviews identified as positive, 89.87% are correct, 10.13% are identified as positive are incorrect. Furthermore, the obtained F1 was 0.72. Novielli et al. used SentiStrenght, a similar lexicon-based classifier as our tool, and obtained with stack-overflow corpus F1 = 0.9 in the positive class. However, with other corpus the result was F1 = 0.65 in the positive class.

4.2.4. Fake review analysis

As our last contribution to the state-of-the-art, we analyse the impact of fake reviews in the previously described experiments. Following the application of the framework, we proceeded to search and eliminate possible fake reviews based on the similarity between several reviews. After calculating the similarity between the 100,000 reviews, using a similarity threshold of 0.85, a total of 1328 reviews were detected as supposed fake reviews in a total of 2200 pairs of reviews. We tested other thresholds and, for this experiment and with this kind of corpus, this is the best threshold. The higher the threshold, the lower the recall. The lower threshold, the lower the accuracy. The results are summarized in Table 8, grouped by three different categories according to the type of duplicate review detected: duplicates from different user identifier (userids) on the same product; duplicates from the same userid on different products; and, duplicates from different userids on different products. As can be observed, the most numerous types correspond to the duplicated reviews posted in different products by different userids, which could be indicative of an intent to introduce the most difficult type of fake review to detect.

Table 10
Changes of the product rating positions due to the elimination of fake reviews.

	By star score	By sentiment value
Positive change	22	31
Negative change	18	16

Table 11
Confusion matrix to evaluate fake review decision tool.

	Real fake review	Real no fake review
Predicted fake review	747	581
Predicted no fake review	81	562

Table 12
Performance of fake review decision tool.

	Accuracy	Recall	F-measure
Fake review	0.5625	0.9021	0.6929
No fake review	0.8740	0.4917	0.6293

Table 13
Confusion matrix to evaluate fake review decision tool.

	Real fake review	Real no fake review	Total
Predicted fake review	668	113	781
Predicted no fake review	81	1109	1190

Table 14
Performance of fake review decision tool.

	Accuracy	Recall	F-measure
Fake review	0.8553	0.8919	0.8732
No fake review	0.9319	0.9075	0.9196

Taking the 1328 reviews that the framework considers as fake reviews, in Table 9, we analyse the most harmful ones that must be spammed: those fake positive reviews for bad average-rated products or fake negative reviews for good average-rated products. Although these fake positive reviews for well-rated products (similarly those fake bad reviews for bad rated products) will not affect the decision-making process, they will abnormally increase/decrease the rating of the product. This information can be used as a key indicator of the likely detection of fake reviews.

After eliminating the fake reviews, the changes in the rating positions (e.g. very bad; bad; regular; good; very good) are analysed. For example, a negative change from “good” to “regular” or from “regular” to “bad”; or a positive change from “regular” to “good”. This analysis is done using the star score and the sentiment value (see Table 10). See (Tables 11–14.)

As a conclusion, these results prove that fake reviews affect the rating of some products and can influence the opinion of the public. Removing them will help to provide more reliable information.

4.2.5. Performance of fake-review detection tool

To analyse the performance of the used fake review detection tool, we use the accuracy, recall and F1 measures. Manually we evaluate each supposed fake review. A review is considered fake if there is a similar fake review of a different product or from a different user.

In our experimentation, we selected 1971 reviews with a high degree of similarities (they have a threshold greater than 0.80). 1328 were predicted as fake reviews, 747 were predicted correctly. 828 were manually considered fake reviews. 81 were not predicted as fake

Table 15
Classification of the main fake review detection methods and their results.

Method. reference	Classification	Subclassification	Dataset	Accuracy
Opinion mining using decision tree based feature selection through manhattan hierarchical cluster measure (Jotheeswaran & Kumaraswamy, 2013	Supervised	Decision tree	IMDb movie	75%
A rule-based approach to emotion cause detection for Chinese micro-blogs (Gao et al., 2015)	Supervised	Rule-based approach	Chinese micro-blog	65%
Machine learning models to classify each review as spam or non-spam using logistic regression. detection of duplicate reviews (Jindal & Liu, 2007; Jindal & Liu, 2008)	Supervised	Logistic Regression	Amazon	78%
Extended SAGE, sparse additive generative, model (Li et al., 2014)	Supervised	Bayesian approach	Hotel, Restaurant and Doctor Reviews, Turker set	65%
Automated classifier performance for three approaches based on nested 5-fold cross-validation experiments. Linguistic n-gram feature based classification approach (Ott et al., 2011)	Supervised	Naïve Bayes and SVM	Hotel reviews through Amazon Mechanical Turk (AMT)	89.8%
SVM 5-fold cross validation classification results across behavioural and bigram features (Mukherjee, Venkataraman, et al., 2013)	Supervised	SVM	Yelp's real-life data	86.1%
Supervised learning using the classification approach random forest and labelling reviews in spam or non-spam classes using sentiment analysis on comment data review (Saumya & Singh, 2018)	Supervised	Random Forest	Amazon Product Reviews	91%
Spamming Behavior regression model, SBM (Lim et al., 2010)	Supervised	Lineal Regression	Amazon Product Reviews	–
Modeling reviewers and their occurrence in bursts as a Markov Random Field, MRF, and employing the Loopy Belief Propagation, LBP, method to infer spammers in the graph (Fei et al., 2013)	Supervised	LBP method	Amazon Product Reviews	77.6%
Upgrading feature-based opinion mining model on vietnamese product reviews (Ha et al., 2011)	Clustering	K-nearest neighbors (KNN)	Vietnamese mobile phone, product review	72%
An unsupervised twice-clustering based product features categorization (Jia et al., 2010)	Clustering	Twice-clustering	Product reviews (360buy.com)	66%
Unified Framework for detecting author spamicity by modeling review deviation using an aspect-based review deviation dimension and a set of abnormity signals from a review deviation angle (Liu & Pang, 2018)	Clustering	Aspect-based review/ latent content deviation	Amazon Product Reviews	78.2%
Group spam rank, GSRank: a novel relation-base approach to rank candidate groups based on likelihoods, using frequent itemset mining method (Mukherjee et al., 2012.	Clustering		Amazon Product Reviews	95%
The author spamicity model, ASM: exploits observed reviewing behaviors to detect opinion spammers in an unsupervised bayesian inference framework based on clustering (Mukherjee, Kumar, et al., 2013)	Clustering		Amazon Product Reviews	77.4%
Textual similarity investigation by constructing time series of reviews (Heydari et al., 2016)	Clustering		Amazon Product Reviews	86%
FraudEagle: Exploits the network effects to automatically detect fraudulent users and fake reviews in online review networks (Akoglu et al., 2013)	Graph based		Entertainment software product (app) reviews anonymous online app store	–
Introduce a new measure NFS, Network Footprint Score, that quantifies the likelihood of products being spam campaign targets devising GroupStrainer to cluster spammers on a 2-hop subgraph induced by top ranking products (Ye & Akoglu, 2015)	Graph based		Amazon, iTunes	100%
Our proposal: Fake review detection framework, FRDF (Kauffmann et al)	Clustering	Cosine similarity measure	Amazon Product Reviews	85.5%

reviews. It produces an accuracy of 56.25% and a recall of 87.40%. The following tables show the confusion matrix, the accuracy, and recall of fake and no-fake reviews.

Additionally, we analysed the erroneously predicted fake reviews and observed that a lot of reviews with less than 3 significant words were wrongly predicted. We proceeded to consider a review as fake if it has a close similarity to other reviews and has more than 3 significant words that are the same. Although only 781 were predicted as fake reviews, the results were better because 668 were predicted correctly and only 113 predicted incorrectly. These results are shown below.

Consequently, applying this improvement, this tool with this corpus achieves 85.53% accuracy in finding the fake review. This means, out of 781 reviews identified as fake reviews, 85,53% were correct, and 14,47% identified as fake review were incorrect.

Table 15 shows the evaluation results of our proposal and the main approaches described in Section 2.2, as well as the descriptors for classification, sub-classification, dataset and accuracy.

We can see that the SVM based methods obtain very good results. We can highlight the methods based on clustering (Mukherjee et al., 2012) and graph-based (Ye & Akoglu, 2015) that achieve the best results. Our proposal based on the measurement of the cosine, which can be classified into clustering techniques, is capable of obtaining an excellent result. Nevertheless, as already mentioned, the main advantage of our proposal is the flexibility and modularity, which allows us to use any method of fake-review detection as presented in Table 15.

4.3. Theoretical and practical contributions

After the experimentation that has been carried out, we can summarize the following theoretical and practical contributions of this work:

1. We have proposed a modular framework that deals with the mining of the textual information included in user reviews, in order to assist marketing managers.
2. This framework considers the inclusion of Sentiment Analysis and Fake Review Detectors tools a must to achieve an optimal analysis of the textual reviews. This modular framework also considers it important to facilitate the link between different tools, as well as the coordinated running between them, given the increasing work in these areas. The coordinated analysis is performed through the dashboards that will allow questions that arise in the background section to be answered (e.g. “what is the most interesting product for the market from the users' point of view?”).
3. We have applied this framework on a corpus of reviews of tech products (e.g. mobile phones) extracted from Amazon, which provides marketing managers with a dashboard showing the rank brand image according to the sentiment expressed by consumers. The dashboard provides the best products based on price, and sentiment values about brand image linked to product prices. For example, if the percentage of negative comments is higher in the expensive products than in the cheap ones. The framework is open to finding other types of relationships between the sentiment value, the price of the product and the star score, or helping to find the best purchase options.
4. From this study, we have analysed the correlation between the score of stars and the result of the sentiment analysis, finding a difference between the rating expressed in stars and opinions. This fact shows that it is important to consider both qualifications to improve product evaluation. The correlation between both scores is measured, and it is shown that often the sentiment value could help discover better products.
5. Regarding the Fake Review task, we have carried out several experiments to detect similar reviews which are classified as possible fake reviews. Evaluations have been carried out with these fake reviews and without them. The conclusion is that they have a strong influence on decision-making from both the company and consumer perspective. This framework discovers other cases of fake reviews and consequently achieves an increase in detection coverage.

5. Conclusions and future work

In this paper, we have proposed a modular framework based on sentiment analysis and the key issue of fake review detection to assist marketing managers and consumers in the decision-making process. The framework provides additional and comparative information mined from consumer reviews and processes them using NLP technology to get sentiment values, a new variable that sheds light on customer behaviour. We have extensively studied the previous work on the issues related to this framework: big data and marketing, sentiment analysis and fake reviews, and the findings are summarized in the background section, in which our contributions to the state-of-the-art are also enumerated.

We have put this framework into practice on a corpus of reviews of tech products extracted from Amazon to rank brand image according to the reviews posted by consumers. We have analysed the correlation between star score and sentiment value finding a difference between both scores, proving that textual reviews contain additional information that is not evident in the star scores. This extra information is important to be considered for reaching a better evaluation of products. In this way, this framework facilitates a comparative analysis that provides answers to important questions both for marketing managers and consumers, such as “What is the best product by price or satisfaction in each category or in an established category?” or “What is the most satisfying product for a given brand?”. Finally, we have analysed the impact of fake reviews in the product rating, by running the experiments with and without the detected fake reviews, in which we have observed that, as expected, they affect the rating.

This paper has some limitations that might restrict the generalization of the results. First, the research framework has been applied only in the retailing platform Amazon. Although this is a very popular marketplace, with great market penetration in Western countries, if the research had been developed in other platforms such as Aliexpress, different results might have been obtained. Second, the sentiment analysis has been carried out with the lexicon Afinn. This lexicon has been widely employed to run this type of analysis. However, it is based on a predetermined rating that was manually constructed by Finn Arup Nielsen. Thus, a different lexicon might provide different results. Finally, the fake reviews detection has considered the cosine similarity measure. Particularly, we set a threshold to consider that a review is a duplicate one. Although we tested other thresholds a greater number of them may have produced different conclusions.

To overcome these limitations, as future work we plan to incorporate additional sentiment analysis tools and fake review detectors in order to comparatively and globally analyse the differences between these tools. Moreover, we plan to analyse more deeply the information in the textual reviews in order to extract positive and negative details evaluated by users, as we highlighted in the experimentation section. To do this effectively, it is vital to detect negations and correctly resolve ambiguity in terms that can express positive and negative sentiments. Finally, the creation of an annotated corpus (according to several annotator opinions) of fake reviews would help the research community to develop tools that assist the fake review detection task.

Acknowledgements

This work was supported in part by the Spanish Research Agency (AEI) and the European Regional Development Fund (FEDER) through the project CloudDriver4Industry under Grant TIN2017-89266-R, in part by the Spanish Ministry of Science, Innovation and Universities through the Project ECLIPSE-UA under Grant RTI2018-094283-B-C32, and in part by the Conselleria de Educación, Investigación, Cultura y Deporte of the Community of Valencia, Spain, within the Program of Support for Research under Project AICO/2017/134 and Project PROMETEO/2018/089.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. *Proceeding LSM '11 proceedings of the workshop on languages in social media* (pp. 30–38). .

Akoglu, L., Chandy, R., & Faloutsos, C. (2013). Opinion fraud detection in online reviews by network effects. *Seventh international AAAI conference on weblogs and social media*.

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics, 24*(1), 1–7.

Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment analysis: Measuring opinions. *Procedia Computer Science, 45*, 808–814. https://doi.org/10.1016/j.procs.2015.03.159.

Brodia, R. (2018). *How to spot fake Amazon reviews.* CNEThttps://www.cnet.com/how-to/spot-fake-amazon-reviews-with-fakespot/.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems, 31*(2), 102–107 (2016).

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion min- ing and sentiment analysis. *Intelligent Systems, IEEE, 28*, 15–21. https://doi.org/10.1109/MIS.2013.30.

Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *Management Information Systems Quarterly, 36*(4), 1165–1188.

Chuang, S. H. (2019). Co-creating social media agility to build strong customer-firm relationships. *Industrial Marketing Management.* https://doi.org/10.1016/j.indmarman.2019.06.012.

Day, M., Wang, C., Chen, C., & Yang, S. (2017). Exploring review spammers by re- view similarity: A case of fake review in Taiwan. *Proceedings of the third international conference on electronics and software science (ICESS2017)* (pp. 166). Takamatsu, Japan.

De Vries, L., Gensler, S., & Leeflang, P. (2012). Popularity of brand posts on brand Fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing, 26*(2), 83–91.

Demoulin, N. T., & Coussement, K. (2018). Acceptance of text-mining systems: The signaling role of information quality. *Information & Management*. https://doi.org/10.1016/j.im.2018.10.006.

Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. *IEEE 24th international conference on data engineering workshop* (pp. 507–512). Cancun.

Elmurngi, E., & Gherbi, A. (2017). Detecting fake reviews through sentiment analysis using machine learning techniques. *Data analytics 2017: The sixth international conference on data analytics* (pp. 65–72). .

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. *Journal of Business Research, 69*(2), 897–904.

Fan, S., Lau, R. Y., & Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Research, 2*(1), 28–32.

Farooq, S., & Khanday, H. (2016). Spam detection : A review. *IJCAI. proceedings of the twenty-second international joint conference on artificial intelligence. 12. IJCAI. proceedings of the twenty-second international joint conference on artificial intelligence* (pp. 1–8). no. 4.

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting burstiness in reviews for review spammer detection. *Seventh international AAAI conference on weblogs and social media*.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89.

Gao, K., Xu, H., & Wang, J. (2015). A rule-based approach to emotion cause detection for Chinese micro-blogs. *Expert Systems with Applications, 42*(9), 4517–4528.

García-Moya, L., Anaya-Sánchez, H., & Berlanga-Llavori, R. (2013). Retrieving product features and opinions from customer reviews. *Intelligent Systems, IEEE, 28*, 19–27. https://doi.org/10.1109/MIS.2013.37.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report. Vol. 1*Stanford (no. 12).

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems, 26*, 191–209.

Ha, Q., Vu, T., Pham, H., & Luu, C. (2011). An upgrading feature-based opinion mining model on Vietnamese product reviews. *International Conference on Active Media Technology,* 173–185.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science, 17*, 26–32.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentimental analysis for twitter accounts. *Mathematical and Computational Applications, 23*(1), 11. https://doi.org/10.3390/mca23010011.

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th international conference on world wide web* (pp. 507–517). .

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management, 52*(7), 801–812.

Heydari, A., Tavakoli, M., & Salim, N. (2016). Detection of fake opinions using time series. *Expert Systems with Applications, 58*, 83–92.

Heydari, A., Tavakoli, M., Salim, N., & Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems with Applications, 42*(7), 3634–3642. https://doi.org/10.1016/j.eswa.2014.12.029.

Holla, L., & Kavitha, K. (2018). A comparative study on fake review detection tech- niques. *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 5*(4).

Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research, 52*(5), 629–641.

Hung, C., & Lin, H. (2013). Using objective words in SentiWordNet to improve word-of-mouth sentiment classification. *IEEE Intelligent Systems, 28*, 47–54.

Jain, A., & Dandannavar, P. (2016). Application of machine learning techniques to sentiment analysis. *In 2016 2nd International conference on applied and theoretical computing and communication technology (ICATCCT)* (pp. 628–632). IEEE.

Jandail, R. R. S. (2014). A proposed novel approach for sentiment analysis and opinion mining. *International Journal of UbiComp, 5*, 1–10. https://doi.org/10.5121/iju.2014.5201.

Jefferson, C., Liu, H., & Cocea, M. (2017). Fuzzy approach for sentiment analysis. *2017 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1–6). .

Jia, W., Zhang, S., Xia, Y., Zhang, J., & Yu, H. (2010). A novel product features categorize method based on twice clustering. *Proceedings of the international conference on web information systems and mining. Vol. 1. Proceedings of the international conference on web information systems and mining* (pp. 281–284). IEEE.

Jiang, M., Cui, P., & Faloutsos, C. (2016). Suspicious behavior detection: Current trends and future directions. *IEEE Intelligent Systems, 31*(1), 31–39.

Jindal, N., & Liu, B. (2007). Analyzing and detecting review spam. *Proceedings - IEEE international conference on data mining, ICDM* (pp. 547–552). .

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *Proceedings of the 2008 international conference on web search and data mining* (pp. 219–230). ACM.

Jotheeswaran, J., & Kumaraswamy, Y. (2013). Opinion mining using decision tree based feature selection through Manhattan hierarchical cluster measure. *Journal of Theoretical and Applied Information Technology, 58*(1), 72–80.

Kumar, A., Shankar, R., & Aljohani, N. (2019). A big data driven framework for demand-driven forecasting with effects of marketing-mix variables. *Industrial Marketing Management.* https://doi.org/10.1016/j.indmarman.2019.05.003.

Lau, R., Liao, S., Kwok, R., Xu, K., Xia, Y., & Li, Y. (2011). Text mining and probabilistic language modeling for online review spam detection. *ACM Transactions on Management Information Systems (TMIS), 2*(4), 25.

Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). *Learning to Identify Review Spam.* IJCAI International Joint Conference on Artificial Intelligence.

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. *Proceedings of the 52nd annual meeting of the association for computational linguistics. vol. 1. Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 1566–1576).

Lim, E., Nguyen, V., Jindal, N., Liu, B., & Lauw, H. (2010). Detecting product review spammers using rating behaviors. *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 939–948). ACM.

Liske, D. (2018). Tidy sentiment analysis in R. https://www.datacamp.com/community/tutorials/sentiment-analysis-R.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1–167.

Liu, X., Burns, A. C., & Hou, Y. (2017). An Investigation of Brand-Related User-Generated Content on Twitter. *Journal of Advertising, 46*(2), 236–247. https://doi.org/10.1080/00913367.2017.1297.

Liu, Y., & Pang, B. (2018). A unified framework for detecting author Spamicity by modeling review deviation. *Expert System with Applications, 112*, 148–155.

Lytras, M., Raghavan, V., & Damiani, E. (2017). Cognitive computing and big data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines. *International Journal on Semantic Web and Information Systems, 13*(1), 1–10.

Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150). .

Mahalakshmi, R. (2017). A study on detecting opinion spam concerning the issues, challenges and opportunities. *IJCRT, 5*(4).

Manning, C., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval.* Cambridge University Press (ISBN-13 978-0-521-86571-5).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). *"The Stanford CoreNLP natural language processing toolkit" proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations.* 55–60. https://doi.org/10.3115/v1/P14-5010.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information & Management, 54*(6), 771–785.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Spotting opinion spammers using behavioral footprints. *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 632–640). .

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on world wide web* (pp. 191–200). ACM.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What yelp fake review filter might be doing? *Proceedings of the international conference on web and social Media* (pp. 409–418). .

Neethu, M., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learn- ing techniques. *4th International Conference on Computing, Communications and Networking Technologies (ICCCNT),* 1–5. https://doi.org/10.1109/ICCCNT.2013.6726818.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science, 31*(3), 521–543.

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement charac- teristics on online review helpfulness: A text regression model. *Decision Support Systems, 61*, 47–58.

Nielsen, F. (2011). A new Anew: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 workshop on 'making sense of Microposts': big things come in small packages 718 in CEUR workshop proceedings* (pp. 93–98). . http://arxiv.org/abs/1103.2903.

Noone, B., & McGuire, K. (2014). Effects of price and user-generated content on consumers' prepurchase evaluations of variably priced services. *Journal of Hospitality and Tourism Research, 38*(4), 562–581.

Novielli, N., Girardi, D., & Lanubile, F. (2018, May). A benchmark study on sentiment analysis for software engineering research. *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)* IEEE (pp. 364-375).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th annual meeting of the association for computational linguistics. vol. 1. Proceedings of the 49th annual meeting of the association for computational linguistics* (pp. 309–319).

Paknejad, S. (2018). *Sentiment classification on Amazon reviews using machine learning approaches.* (Dissertation).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.

Paracchini, P. (2016). Sentiment analysis using the tidytext package. https://rpubs.com/pparacch/236096.

Rahman, K., & Khamparia, A. (2016). Techniques, applications and challenges of opinion mining. *IJCTA International Journal of Control Theory and Applications, 9*(41), 455–461.

Sathi, A. (2014). *Engaging customers using big data: How marketing analytics are transforming business.* New York: Palgrave Macmillan.

Saumya, S., & Singh, J. (2018). Detection of spam reviews: A sentiment analysis approach. *Csi Transactions on ICT, 6*(2), 137–148.

Shivagangadhar, K., Sagar, H., Sathyan, S., & Vanipriya, C. (2015). Fraud detection in online reviews using machine learning techniques. *International Journal of Computational Engineering Research (IJCER), 5*(5), 52–56.

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software, 1*(3), 37.

Sindhu, C., Vyas, D., & Pradyoth, K. (2017). Sentiment analysis based product rating using textual reviews. *2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), 2*, 727–731. https://doi.org/10.1109/ICECA.2017.8212762.

Singla, Z., Randhawa, S., & Jain, S. (2017). *Sentiment analysis of customer product.*

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, 70*, 263–286.

Sun, T., Youn, S., Wu, G., & Kuntaraporn, M. (2006). Online word-of-mouth: An exploration of its antecedents and consequences. *Journal of Computer-Mediated Communication, 11*(4), 1104–1127.

Tavakoli, M., Heydari, A., Ismail, Z., & Salim, N. (2016). A framework for review spam detection research. *World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering, 10*(1), 67–71.

Team, R. C. (2013). *R: A language and environment for statistical computing.*

Tsang, A. S. L., & Prendergast, G. (2009). Is a star worth a thousand words? The interplay between product-review texts and rating valences. *European Journal of Marketing, 4*(11/12), 1269–1280.

Wahyuni, E., & Djunaidy, A. (2016). Fake review detection from a product review using modified method of iterative computation framework. *MATEC Web of Conferences, 58*, 03003. https://doi.org/10.1051/matecconf/20165803003.

Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). Towards felicitous decision making: An overview on challenges and trends of big data. *Information Sciences, 367*, 747–765.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environ- ments. *Journal of Marketing, 80*(6), 97–121.

Wu, Z., Wang, Y., Wu, J., Cao, J., & Zhang, L. (2015). Spammers detection from product reviews: A hybrid model. *2015 IEEE International Conference on Data Mining,* 1039–1044.

Ye, J., & Akoglu, L. (2015). Discovering opinion spammer groups by network footprints. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 267–282). .

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Ex- tracting sentiments about a given topic using natural language processing techniques. *Proceedings of the 3rd IEEE international conference on data mining* (pp. 427–434). .

Zhang, K., & Katona, Z. (2012). Contextual advertising. *Marketing Science, 31*(6), 980–994.

Zhang, L., Wu, Z., & Cao, J. (2018). Detecting spammer groups from product reviews: A partially supervised learning model. *IEEE Access, 6*, 2559–2568.