Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Lean Startup Management

MGT1022

Assignment 2

Reviewing Research Papers -
Market Plan Including Digital and Viral Marketing

Slot : TE2

Kulvir Singh - 19BCE2074
Gurtavrein Singh – 19BCE2101

Market Plan Including Digital and Viral Marketing

Paper 1 : Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics

Impact Factor : 9.269

Year : 2021

Methodology Adopted :

The paper discusses and explains the fact that since the beginning of the digital age, there has been an influence of technology in almost all domains of the society. Digital Marketing is one such domain which has been vastly influenced by technology especially the use of Data Sciences.
The influence of data on the digital marketing ecosystem. The way to boost the Return on Investment (ROI) was explored. The key methods used to explore the possibility of data science applications in digital marketing were analyzed. The key points where DS (Data Science) can be applied are Search Engine Optimization, Search Engine Marketing, programmatic advertisements, Social Media Marketing. The key domains of data science applied on the above mentioned points are data collection, data tuning, artificial intelligence, information sciences and machine learning techniques.

Results Obtained :

The paper was able to classify the results obtained dataset analysis into data sources such that data science can be applied on it with respect to digital marketing. The dataset was tuned and classified into 4 types of data. Transactional Data, Non-Transactional Data, Operational Data, Online Data (sources).

The paper was also able to identify the various machine learning and data science paradigms that could be applied on the classified dataset. The programming models basically included Ensemble Models(EM), Deep Learning Neural Networks, Machine Vision, Natural Language Processing.

Critical Analysis of the results :

The analysis of the result comprises of the explanation of the 4 types of data and the 4 types of paradigms that can applied on the data.
Transactional data simply specified the information regarding the sales and payments of the business. Non-transactional data showed the demographic and lifestyle of the customers. Operational data was classified as data based on strategies used for marketing. Online data was basically the data scrapped from the web. The collected data then underwent the 4 above mentioned techniques and it was seen that it provided statistical result for improving the digital marketing of a company. Therefore the paper concluded its finding by stating that the quality and quantity of digital marketing can be modified and determined by the statistical analysis of results provided by DS.

Bibliography:

Saura, J. R. (2021). Using data sciences in digital marketing: Framework, methods, and performance metrics. Journal of Innovation & Knowledge, 6(2), 92-102.

Paper 2 : Digital marketing strategies, online reviews and hotel performance

Impact Factor : 9.237

Year : 2018

Methodology Adopted :

The research was conducted in various countries such as Flanders, Belgium: Antwerp, Bruges, Ghent, Mechelen and Leuven. There were a huge percentage of 1 to 3 star hotels and a few 4 to 5 star hotels that were included as a part of the survey. A review survey was conducted using a questionnaire pamphlet which was given to the guests at the hotels. The questionnaire consisted of various questions such as how did you find the hotel, did you come across any online or digital ads of the same etc. A complex survey and review were also done for the websites for each of the hotels.

Results Obtained :

The result from the surveying and the review of the online websites were compiled and tabulated. The results were tabulated according to the parameters such as relevancy of the ads and websites, proper campaign of ads, reachability and visibility and some small factors such as brand value and luxury. This provided a holistic view on the approach of marketing strategy of each hotel and it enabled the researchers to gauge the influence of digital and online marketing and its benefits and importance on the hotel industry from this case study.

Critical Analysis of the results :

For analyzing the result, the researchers created a parameter called the confidence interval. This range of classifies the result tabulated into a single entity which can be associated to all the reviewed hotels. The range showcases the effectiveness of the digital marketing scheme of the hotels with respect to the number of guests attracted by the marketing strategy and also the willingness of the guest to return back and share the digital ads to other people. The researchers concluded that the effective use of digital marketing strategies has brought about a rise in the attraction level of the hotels leading to an increase in the business of the same.

Bibliography:

De Pelsmacker, P., Van Tilburg, S., & Holthof, C. (2018). Digital marketing strategies, online reviews and hotel performance. International Journal of Hospitality Management, 72, 47-55.

Paper 3 : A Study on the Impact of Social Media Marketing Trends on Digital Marketing

Impact Factor : 3.122

Year : 2018

Methodology Adopted :

A comprehensive literature review was conducted by the researchers on the topic of underlying usage of social media, the attitude towards social media marketing messages, and the effectiveness of messages about online shopping value. The researchers also reviewed documents which pertained to explore how to engage a series of customers and attract an audience on social media and similar marketing audience. The researchers then brief us about the social networking websites and how they can be used to promote a business and become a primary source of advertisement as well as the primary shop setup which is free of cost unless an additional promotion is applied and given to the social networking site. The researchers also explain how mobile phones will become the eventual window of interaction between customers and the business. Lastly, they discuss about the trends in social media marketing and the general impact of social media marketing in the entire market ecosystem.

Results Obtained :

The surveying led to multiple valuable conclusions. Social networking is used by about 76% of businesses to achieve their marketing objectives. Business retailers experience about 133% increase in revenues after marketing their business in the mobile market that promotes social media marketing value for their business. 40% of online shoppers from the US use the Smartphone for in-store shopping. About 71% of the consumers respond according to the feedback and recommendation of social users regarding a particular brand. Consumer reviews are regarded by shoppers as trustworthy than the marketing promotion coming directly from the brand site. The majority of successful brands have a social media page to widen their marketing coverage of making their brand more accessible among social media users. It was also found that businesses would be willing to switch to image-centric content for social media marketing along with social integration to email marketing as it was highly beneficial for attracting the attention of the public.

Critical Analysis of the results :

The research started with the aim to analyze the different issues related to digital marketing. Based on the discussion it has been found that in the case of digital marketing the most important aspect is to connect with the users. The ladder of engagement has shown the approaches to attach to the customers. The study has also revealed that to utilize digital marketing effectively, the companies are required to design an effective platform. With the example of Interest, the effectiveness of a social media platform has been discussed. The current trends in digital marketing have also been discussed in the research.

Bibliography:

Ibrahim, S. S., & Ganeshbabu, P. (2018). A Study on the Impact of Social Media Marketing Trends on Digital Marketing. Shanlax International Journal of Management, 6(1), 120-125.

Plagiarism Report :

Plagiarism Detector.net

Date: April, 06 2022

PLAGIARISM SCAN REPORT

12%	88%	892	5600
Plagiarised	Unique	Words	Characters

Excluded Url : None

Content Checked For Plagiarism

The paper discusses and explains the fact that since the beginning of the digital age, there has been an influence of technology in almost all domains of the society. Digital Marketing is one such domain which has been vastly influenced by technology especially the use of Data Sciences. The influence of data on the digital marketing ecosystem. The way to boost the Return on Investment (ROI) was explored. The key methods used to explore the possibility of data science applications in digital marketing were analyzed. The key points where DS (Data Science) can be applied are Search Engine Optimization, Search Engine Marketing, programmatic advertisements, Social Media Marketing. The key domains of data science applied on the above mentioned points are data collection, data tuning, artificial intelligence, information sciences and machine learning techniques. The paper was able to classify the results obtained dataset analysis into data sources such that data science can be applied on it with respect to digital marketing. The dataset was tuned and classified into 4 types of data. Transactional Data, Non-Transactional Data, Operational Data, Online Data (sources). The paper was also able to identify the various machine learning and data science paradigms that could be applied on the classified dataset. The programming models basically included Ensemble Models(EM), Deep Learning Neural Networks, Machine Vision, Natural Language Processing. The analysis of the result comprises of the explanation of the 4 types of data and the 4 types of paradigms that can applied on the data. Transactional data simply specified the information regarding the sales and payments of the business. Non-transactional data showed the demographic and lifestyle of the customers. Operational data was classified as data based on strategies used for marketing. Online data was basically the data scrapped from the web. The collected data then underwent the 4 above mentioned techniques and it was seen that it provided statistical result for improving the digital marketing of a company. Therefore the paper concluded its finding by stating that the quality and quantity of digital marketing can be modified and determined by the statistical analysis of results provided by DS. The research was conducted in various countries such as Flanders, Belgium: Antwerp, Bruges, Ghent, Mechelen and Leuven. There were a huge percentage of 1 to 3 star hotels and a few 4 to 5 star hotels that were included as a part of the survey. A review survey was conducted using a questionnaire pamphlet which was given to the guests at the hotels. The questionnaire consisted of various questions such as how did you find the hotel, did you come across any online or digital ads of the same etc. A complex survey and review were also done for the websites for each of the hotels The result from the surveying and the review of the online websites were compiled and tabulated. The results were tabulated according to the parameters such as relevancy of the ads and websites, proper campaign of ads, reachability and visibility and some small factors such as brand value and luxury. This provided a holistic view on the approach of marketing strategy of each hotel and it enabled the researchers to gauge the influence of digital and online marketing and its benefits and importance on the hotel industry from this case study. For analyzing the result, the researchers created a parameter called the confidence interval. This range of classifies the result tabulated into a single entity which can be associated to all the reviewed hotels. The range showcases the effectiveness of the digital marketing scheme of the hotels with respect to the number of guests attracted by the marketing strategy and also the willingness of the guest to return back and share the digital ads to other people. The researchers concluded that the effective use of digital marketing strategies has brought about a rise in the attraction level of the hotels leading to an increase in the business of the same. A comprehensive literature review was conducted by the researchers on the topic of underlying usage of social media, the attitude towards social media marketing messages, and the effectiveness of messages about online shopping value. The researchers also reviewed documents which pertained to explore how to engage a series of customers and attract an audience on social media and similar marketing audience. The researchers then brief us about the social networking websites and how they can be used to promote a business and become a primary source of advertisement as well as the primary shop setup which is free of cost unless an additional promotion is applied and given to the social networking site. The researchers also explain how mobile phones will become the eventual window of interaction between customers and the business. Lastly, they discuss about the trends in social media marketing and the general impact of social media marketing in the entire market ecosystem. The surveying led to multiple valuable conclusions. Social networking is used by about 76% of businesses to achieve their marketing objectives. Business retailers experience about 133% increase in revenues after marketing their business in the mobile market that promotes social media marketing value for their business. It was also found that businesses would be willing to switch to image-centric content for social media marketing along with social integration to email marketing as it was highly beneficial for attracting the attention of the public.

6% Plagiarized

Using Data Sciences in Digital Marketing: Framework, methods, and ...

https://zh.art1lib.com/book/83776376/d7f685

6% Plagiarized

Mar 17, 2014 · Business retailers experience about 133% increase in revenues after marketing their business in the mobile market that promotes social media marketing value for their business. 40% of online shoppers from the US use the Smartphone for in-store shopping.

https://www.socialmediatoday.com/content/impact-social-media-marketing-trends-digital-marketing/

Plagiarism Detector.net

Paper 1: Examining the relationship between social media analytics practices and business performance in the Indian retail industry: The mediating role of customer engagement

International Journal of Information Management

Impact Factor: 14.098

Methodology Adopted:

This paper highlights link between social media analytics practice (SMAP), customer engagement (CE), and business performance (BP). In order to gain a better understanding of the relationship between SMAP and BP and the mediation role of CE in that process, author has conducted a large-scale survey among senior and mid-level managers. Specifically, a structured closed-ended questionnaire was administered to managers and management consultants' country-wide and gathered usable responses from 281 respondents holding positions. The objective of this research, was to empirically investigate and produce knowledge

about the nature of the relationship that exists between customer engagement and business

performance. They have used SMAP as the tool for investigation. Four research questions have

been identified in their research as follows:

(1) Does strategic use of SMAP have a positive relationship with customer engagement?

(2) Does customer engagement have a positive relationship with business performance?

(3) Does strategic use of SMAP have a positive relationship with business performance?

(4) Does customer engagement have a mediating effect on the relationship between

SMAP and Business performance?

Results Obtained:

The paper was able to concluded that there exists a positive relationship between SMAP and BP in which CE plays a key mediation role. It examines four research questions and to address these questions a comprehensive model was developed and tested using Structural Equation Modelling (SEM) analysis. This research paper provides empirical justifications for the existence of a causal relationship between SMAP, CE, and BP. It provides empirical evidence to support the theoretical and prescriptive statements in the literature.

Critical Analysis of the results:

This research paper also has limitations which affect its generalisation. Firstly, although they have considered widely accepted items of SMAP, CE and BP derived from the literature there is the possibility that they may not have included in the research some of the items which are less common in the literature. Moreover, the findings of their research relate solely to Indian retail and IT industries. As such it may not constitute a sufficient basis for generalisation

Bibliography:

Garg, P., Gupta, B., Dzever, S., Sivarajah, U., & Kumar, V. (2020). Examining the relationship between social media analytics practices and business performance in the Indian retail and IT industries: The mediation role of customer engagement. *International journal of information management*, *52*, 102069.

Paper 2: Social media metrics and analytics in marketing – S3M: A mapping literature review

International Journal of Information Management

Impact Factor: 14.098

Methodology Adopted:

The purpose of this paper was to present a mapping literature review and a classification for research articles regarding social media metrics and analytics in marketing. This paper covers 52 articles from peer review journals and international conferences, from 2010 to 2016. These 52 articles are classified in 5 distinct categories based on their: methodology of research, type of analysis, field of study, marketing objectives and social media type/ platform used. The findings of the study revealed which is the most used subcategory for each classification, trends and tendencies.

Results Obtained:

This paper found that there is a peak on publications in 2013 followed by a decrease the next two years. An important finding is that 2016 represents a small but constant increase in the number of publications, showing an overall increase of interest in social media marketing analysis. Trends show tourism industry, Facebook and Twitter as well as consumer-centric marketing to be the dominant categories, platforms and concepts behind social media marketing strategies. On the other hand, though, these trends may bring to the surface gaps in other fields that need attention and research.

Critical Analysis of the results:

This paper was conducted with keywords such as “social media marketing” and not separately for each marketing objective (e.g. branding, engagement, etc.). This fact, limited the number of the articles. Future studies must approach the S3M topic, by searching (and adding as keywords) every field; platform and marketing objective; separately. This type of search will lead to diverse studies; focusing on a specific direction. Our proposed S3M typology framework should trigger future research enabling the incorporation of further criteria.

Bibliography:

Misirlis, N., & Vlachopoulou, M. (2018). Social media metrics and analytics in marketing–S3M: A mapping literature review. *International Journal of Information Management*, *38*(1), 270-276.

Paper 3: Active Viral Marketing: Incorporating Continuous Active Seeding Efforts into the Diffusion Model

Impact Factor: 9.654

Methodology Adopted:

This paper proposes a new diffusion model, named Active Viral Marketing, which better fits real-world marketing scenarios, where adoption of products relies on continuous active promotion efforts by the marketer. The paper further proposes a set of heuristics to schedule the marketing attempts. The main idea behind these heuristics is to utilize the information on the dynamic adoption-states of neighbour nodes, in addition to the static social network topology, when choosing the next node to seed.

Results Obtained:

This method proposed a new information diffusion model, named Active Viral Marketing (AVM), in which agents, e.g., sales representative of a company, communicate with network users, e.g., potential clients, and offer them a new product or service. The probability that a user accepts such an offer is based on the previous adoption rate of his/her friends, as well as his/her own tendency toward the product.

Critical Analysis of the results:

The proposed method is mainly applicable to products that have a viral characteristic. These are products or services where a substantial part of the purchasing decision is based on social influence. In products or services for which the social forces are significantly less influential, it might be better to use the existing state-ofthe-art methods of selecting nodes based on the network's topological properties.

Bibliography:

Sela, A., Goldenberg, D., Ben-Gal, I., & Shmueli, E. (2018). Active viral marketing: Incorporating continuous active seeding efforts into the diffusion model. *Expert Systems with Applications*, *107*, 45-60.

A Study on the Impact of Social Media Marketing Trends on Digital Marketing

Dr.S.A. Shamsudeen Ibrahim,
M.Com., M.Phil, B.Ed., PGDCA, PGDCM., Ph.D.,
Assistant Professor of Commerce
Dr.Zakir Husain College, Ilayangudi

P.Ganeshbabu, M.C.A., M.Phil, TNSET, MBA,
Head – Assistant Professor of Computer Science
Senthamarai College of Arts and Science, Madurai

OPEN ACCESS

Volume: 6

Special Issue: 1

Month: October

Year: 2018

ISSN: 2321-4643

Impact Factor: 3.122

Citation:
Shamsudeen Ibrahim, S. A. &., and P. Ganeshbabu. "A Study on Impact of Social Media Marketing Trends on Digital Marketing." *Shanlax International Journal of Management*, vol. 6, no. S1, 2018, pp. 120–125

DOI:
https://doi.org/10.5281/zenodo.1461321

Abstract

Growing popularity of social media compelled the marketers to think about this media along with traditional functional areas of marketing. Social media is based primarily on the internet or cellular phone-based applications and tools to share information among people. The number of social media user's is more than the population of some of the countries today. Impact of social media on marketing can be judged by comparing marketing before social media and marketing after the introduction of social media and the type of technologies used in social media

Keywords: Marketing before social media, the evolution of social media, social media today, and web technology, the impact of social media on marketing, concerns, and criticism of social media.

Introduction

Social media marketing is the use of social media platforms and websites to promote a product or service. Although the terms e-marketing and digital marketing are still dominant in academia, social media marketing is becoming more popular for both practitioners and researchers. Most social media platforms have built-in data analytics tools, which enable companies to track the progress, success, and engagement of ad campaigns. Companies address a range of stakeholders through social media marketing, including current and potential customers, current and potential employees, journalists, bloggers, and the general public. On a strategic level, social media marketing includes the management of a marketing campaign, governance, setting the scope (e.g., more active or passive use) and the establishment of a firm's desired social media "culture" and "tone.". When using social media marketing, firms can allow customers and Internet users to post user-generated content (e.g., online comments, product reviews, etc.), also known as "earned media," rather than use marketer-prepared advertising copy.

The Objective of the Study

- To study Social Media Marketing
- To study Digital Marketing
- To study Digital Marketing Trends

Methodology

Research Type: Descriptive Type

Data Collection: Secondary data Collected through leading Journals, reviews, chapter Books.

Review of Literature

Chung and Austria (2010) researched with objectives to find out, what gratifications are underlying the usage of social media, the attitudes towards social media marketing messages, and the effectiveness of messages about online shopping value. The base was taken on the Uses and Gratification theory (Katz, Bluner & Gurevitch, 1974 and as enhanced by Ruggiero, 2000), to investigate consumer gratification in social media usage. Online shopping value was examined in a relationship with social media marketing messages. For social media gratifications, entertainment, information, and interaction were taken as exogenous variables. Attitude towards social media marketing messages and online shopping values were the endogenous variables.

Minton, Lee, Orth, Kim, and Kahle (2012) did this very interesting research on sustainable marketing and social media, involving cross-culture populations (subjects) to analyze the motives for sustainable behaviors. South Koreans are representing collectivist culture and USA, Germany being more of individualistic culture, were studied based on their usage of Face book and Twitter about motives for sustainable behaviors. Using Kelman's (1958) functional motives as a basic theoretical foundation, online survey method was used to cover the subjects belonging to different cultures. The conceptual model for this study tried to analyze how functional motives (responsibility, involvement, and internalization) influence the sustainable behaviors such as recycling behaviors, organic food purchase, green transport use, anti-materialistic views, and charity.

Vinerean, Cetina, Dumitrescu, and Tichindelean (2013) did this exploratory research based on primary data using university students in Romania to explore how to engage with different types of an audience on social media marketing platforms (based on their online behavioral aspects), to maximize the effect of online marketing strategy. A linear model was examined to find out how different predictors related to online users and social networking sites, have a positive impact on audiences perceptions of online advertisement.

The Platform of Marketing

Social networking websites

Social networking websites allow individuals, businesses and other organizations to interact with one another and build relationships and communities online. When companies join these social channels, consumers can interact with them directly. That interaction can be more personal to users than traditional methods of outbound marketing and advertising. Social networking sites act as word of mouth or more precisely, e-word of mouth. The Internet's ability to reach billions across the globe has given an online word of mouth a powerful voice and far reach. The ability to rapidly change buying patterns and product or service acquisition and activity to a growing number of consumers is defined as an influence network. Social networking sites and blogs allow followers to "retreat" or "repost" comments made by others about a product being promoted, which occurs quite frequently on some social media sites. By repeating the message, the user's connections can see the message, therefore reaching more people. Because the information about the product is being put out there and is getting repeated, more traffic is brought to the product/company.

Social networking websites are based on building virtual communities that allow consumers to express their needs, wants and values, online. Social media marketing then connects these consumers and audiences to businesses that share the same needs, wants, and values. Through social networking sites, companies can keep in touch with individual followers. This personal interaction can instill a feeling of loyalty into followers and potential customers. Also, by choosing whom to follow on these sites, products can reach a very narrow target audience. Social networking sites also include much information about what products and services prospective clients might be interested in. Through the use of new semantic analysis technologies, marketers can detect buying signals, such as content shared by people and questions posted online. An understanding of buying signals can help sales people target relevant prospects, and marketers run micro-targeted campaigns.

In 2014, over 80% of business executives identified social media as an integral part of their business. Business retailers have seen 133% increases in their revenues from social media marketing.

Mobile phones

More than three billion people in the world are active on the Internet. Over the years, the Internet has continually gained more and more users, jumping from 738 million in 2000 all the way to 3.2 billion in 2015. Roughly 81% of the current population in the United States has some type of social media profile that they engage with frequently. Mobile phone usage is beneficial for social media marketing because mobile phones have social networking capabilities, allowing individuals immediate web browsing and access to social networking sites. Mobile phones have grown at a rapid rate, fundamentally altering the path-to-purchase process by allowing consumers to easily obtain pricing and product information in real time and allowing companies to constantly remind and update their followers. Many companies are now putting QR (Quick Response) codes along with products for individuals to access the company website or online services with their smart phones. Retailers use QR codes to facilitate consumer interaction with brands by linking the code to brand websites, promotions, product information, or any other mobile-enabled content. Also, Real-time bidding use in the mobile advertising industry is high and rising because of its value for on-the-go web browsing. In 2012, Nexage, a provider of real-time bidding in mobile advertising, reported a 37% increase in revenue each month. Adfonic, another mobile advertisement publishing platform, reported an increase of 22 billion ad requests that same year.

Mobile devices have become increasingly popular, where 5.7 billion people are using them worldwide, and this has played a role in the way consumers interact with media and has many further implications for TV ratings, advertising, mobile commerce and more. Mobile media consumption such as mobile audio streaming or mobile video are on the rise – in the United States, more than 100 million users are projected to access online video content via mobile device. Mobile video revenue consists of pay-per-view downloads, advertising, and subscriptions. As of 2013, worldwide mobile phone Internet user penetration was 73.4%. In 2017, figures suggest that more than 90% of Internet users will access online content through their phones.

Impact of Social Media Marketing Trends on Digital Marketing

The growth of social media marketing platforms has become a major part of building social signals that are very important in any SEO digital marketing campaign. Perhaps you are unaware that the emergence of different social media channels offers internet marketers like you a wider marketing opportunities in building brand visibility over the web. How your website ranks on the search engine can make a big impact regarding your customer and lead acquisition and conversion rate for your site. Social media marketing integrated with search engine optimization strategies is effective in building an organic for website traffic. There are different social media marketing trends that will affect the way digital marketers will undertake their search engine optimization campaign to boost their lead generation process and website conversion rates this year.

From the insights of digital marketing experts, here are some of the social media marketing trends that can impact the growth and success of your digital marketing and search engine optimization campaigns. Are you ready to embrace these trends into integrating them to your internet marketing structures?

Investing in social media marketing - A need than a want

Online marketers now view the value of social media marketing for their business from a different perspective. There is a significant explosion in the number of consumers who are using socials as a means of finding products and services that they need. According to prestigious social consumer statistics:

- Social networking is used by about 76% of businesses to achieve their marketing objectives.
- Business retailers experience about 133% increase in revenues after marketing their business in the mobile market that promotes social media marketing value for their business.
- 40% of online shoppers from the US use the Smartphone for in-store shopping.
- About 71% of the consumers respond according to the feedback and recommendation of social users regarding a particular brand.
- Consumer reviews are regarded by shoppers as trustworthy than the marketing promotion coming directly from the brand site.
- The majority of successful brands have a social media page to widen their marketing coverage of making their brand more accessible among social media users.

Among the benefits of using social media channels in promoting a brand include

1. Growing social signals

Social signals can significantly boost your search engine optimization efforts. The more people in the social media community share, like, recommend and talk about your business the more the search engine finds your website relevant thereby increasing the chance of your web pages to acquire a higher position to the search engine results page.

2. Promote company branding and awareness

Social media users can always recommend to their social media circles significant quality of your brand. This can be a good marketing boost to your brand image and in growing the number of people becoming more interested in your brand reputation and to become a follower of your brand.

3. Word of mouth advertising is powerful

Word of mouth advertising tends to have a higher trust rating from the consumers than the product descriptions that your company promotes from your site. Whenever your web page gains more likes and shares from the social media community, the wider your audience reach and influence becomes to your target customers.

Integrating social media into your digital marketing campaign is thus crucial to attaining your marketing goals. From becoming a mere luxurious means of marketing business online, social media marketing becomes an important pillar in SEO with the need of integrating it to digital marketing in an effort of making small to medium businesses at par and competitive with their competitors.

The social advertising trend becoming indispensable in digital marketing

Digital marketers are being lured towards social media advertising due to the trend in the shopping behavior of the consumers. Social media surveys reveal that a big percentage of the consumers spend an average of 37 minutes a day on popular social media channels like Facebook and Twitter and 10% of the internet users are spent on social media sites. Imagine the potential market gain that social media can offer to online marketers. In 2013, about 53% of digital marketers were already positioning their brand in the social media market and by 2014 social advertising investment will continue to grow. If your business is not taking this marketing step to grow your market coverage by now, your competitors are probably taking advanced steps with a better and large market opportunity to play around.

To leverage social advertising to your business advantage, it is essential to implement the following:

- Define measurable goals for your business
- Integrate social advertising into your search engine optimization strategy to optimize your marketing efforts and results.
- Identify your target customer behavior, needs, and activities by using SEO analytics. This will help measure the potential effectiveness on the kind of social advertisement approach to using when engaging your target audience.
- Optimize the landing pages of your website by combining SEO and social media marketing strategies. Don't underestimate the influence of social media buttons in making your landing pages more engaging to your website visitors.

Image-Centric content for social media marketing

Social media users are becoming more engaged in sharing images and liking them. For an internet marketer, this is a good opportunity to grow the social signals that will give their brand a better search rank. Image content can be very enticing among the social media users that offer a good online exposure of a brand. The image-centric content has become one of the social media marketing trends embraced by Atlanta's social media marketing companies that are likewise integrated to their search engine optimization campaigns under the principle that mages are known to boost a brand's exposure to the search engine users.

Social integration to email marketing

Email marketing is viewed by digital marketers as one of the pillars for successful lead conversion. The widespread use of email marketing remains to be prevalent despite the latest trend in digital marketing and marketers are taking the initiative of integrating social media marketing to further strengthen their business lead conversions. By using social media, your leads will find it easier to make a buying decision if they see your brand within their friends' social feeds. Social media marketers usually employ the process of updating their email marketing content in their social media status updates which effective in promoting brand marketing updates.

Conclusion

The study started with the aim to analyze the different issues related to digital marketing. Based on the discussion it has been found that in the case of digital marketing the most important aspect is to connect with the users. The ladder of engagement has shown the approaches to attach to the customers. The study has also revealed that to utilize digital marketing effectively, the companies are required to design an effective platform. With the example of Interest, the effectiveness of a social media platform has been discussed. The current trends in digital marketing have also been discussed in the study. It has shown that in the current context, it has become important to integrate all the systems with that of the digital platform. The transition of a newspaper from the printed version to the online version has been exemplified the current trends of the digitalization.

References

https://en.wikipedia.org/wiki/Social_media_marketing
https://www.socialmediatoday.com/content/impact-social-media-marketing-trends-digital-marketing
www.google.com
Marketing Management

Web Sources

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.434.2756&rep=rep1&type=pdf
https://play.google.com/store/apps/details?id=com.shahabfktech.socialmediamarketing&hl=en_US
http://25razor.com/social-media-experts-information/
http://howchoices.blogspot.com/2017/08/how-to-start-social-media-business.html
https://en.wikipedia.org/wiki/Electronic_word-of-mouth
http://www.mainesocialmedia.com/
http://pluminternet.weebly.com/
https://www.socialmediatoday.com/content/impact-social-media-marketing-trends-digital-marketing

International Journal of Information Management 38 (2018) 270–276

Contents lists available at ScienceDirect

International Journal of Information Management

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

Review

Social media metrics and analytics in marketing – S3M: A mapping literature review

Nikolaos Misirlis*, Maro Vlachopoulou

University of Macedonia, Egnatias 156, Thessaloniki, Greece

ARTICLE INFO

Keywords:
Social media
Social media analytics
Marketing
Literature review
Framework

ABSTRACT

The purpose of this study is to present a mapping literature review and a classification for research articles regarding social media metrics and analytics in marketing. The review covers 52 articles from peer review journals and international conferences, from 2010 to 2016. These 52 articles are classified in 5 distinct categories based on their: *methodology of research, type of analysis, field of study, marketing objectives and social media type/platform used.* The findings of the study reveal which is the most used subcategory for each classification, trends and tendencies. This review provides a base classification for researchers and an editable and continuously augmenting typology for further research in the area.

1. Introduction

Web 2.0 tools and the appearance of social media seem to have redefined the marketing strategy, research and practice, broadening marketing's potential. These potentials go beyond customers' information and expand on commitment and engagement levels. Constantinides and Fountain (2008) define Web 2.0 "as a collection of open-source, interactive and user-controlled online applications expanding the experiences, knowledge and marketing power of the users as participants in business and social process […] supporting the creation of informed users' networks facilitating the flow of ideas and knowledge by allowing the efficient generation, dissemination, sharing and editing/refining of information content".

Social media produce a vast amount of measurable useful data to analysts and marketers whose goal is to monitor and analyze behavioral targeting, brand loyalty and further marketing performance indicators, rendering these data effective. To do that, specific marketing metrics goals need to be clearly defined. Without a specific plan, regarding also the key performance indicators choices, data analysts together with marketers will fail to direct the social media data into useful insights for the companies. For that purpose, firms must precisely raise questions and search answers from social media listening in order to transform data in social media metrics. Social media analysis, therefore, consists of collecting, measuring, evaluating and finally interpreting data (Kaplan & Haenlein, 2010).

Since the first appearance of social media, marketers have noticed the potential of such technology in business (Mangold & Faulds, 2009). Social media can serve as an effective marketing tool in business, valuable for both consumers and companies, offering a wide range of opportunities (Kaplan & Haenlein, 2010). Therefore, social media show an unprecedented increase of use inside business. Even though, understanding social media is a crucial, but not a simple procedure. Several definitions are classified in order to fully explore the dynamics of social media in marketing.

This study presents a complete base for understanding and describing social media metrics and social media analytics related to marketing strategy, policy and research, by reviewing the relevant literature. The objective of this paper is an extensive review of articles related to social media metrics and analytics in marketing, creating a mapping review/systematic map of the relevant material. The primary goal in this article is to create a conceptual classification scheme (named S3M) for the extant literature by using five distinct dimensions/criteria of classification, such as: Methodology of research, Type of analysis, Field of study, Marketing objectives, and Social media types/platforms. As a result, the most used subsectors from each category are identified, featuring the new upcoming trends in social media marketing. The findings of this study are expected to benefit researchers and marketers by helping them to better understand what has been hitherto achieved. It is our primary hope that the proposed framework will serve as a valuable classification system for researchers, academics and practitioners who conduct similar research.

The paper is structured as follows. The next section presents the research methodology we follow. In the following section we present the classification of the literature, providing a discussion section for each category. The final section summarizes our work, offering concluding remarks, future research directions and limitations that rise from our study.

* Corresponding author at: Egnatias 156, Thessaloniki, Greece.
E-mail address: nikosmisirlis@gmail.com (N. Misirlis).

http://dx.doi.org/10.1016/j.ijinfomgt.2017.10.005
Received 24 August 2016; Received in revised form 28 September 2017; Accepted 23 October 2017

2. Research methodology

Social media marketing as a science field is difficult to restrict only in few specific disciplines. This difficulty arises due to the multi-disciplinary nature of the sciences and industry fields involved. Based on our proposal, articles associated to S3M can be found in five types of journals: Marketing and e-Marketing, E-Business and Management, Behavioral sciences, ICT/Information systems and Social media. In order to limit the collection of articles, we take some restrictions into consideration. The articles were initially searched on Internet and academic databases such as Science Direct, Scopus and Emerald. Articles from books and book chapters are excluded from the research. The search returned 101 articles, covering the time span 2011–2016. Of them, 35 were rejected due to lack of compatibility of the content with our research scope. From the 66 remaining relative articles, we excluded 6, for being white papers. From the remaining 60 articles, 52 are scientific articles from peer-review journals and 8 from conferences and proceedings. Each article was reviewed and classified initially into the five above mentioned categories and furthermore in relation with the year of publication. The year distribution can reveal useful outcomes for the research tendencies.

As it is shown in Fig. 1, the research has increased significantly since 2012. This year together with 2014 contribute 8 articles. The pick on publications is noticed during 2013 with 12 articles. 2014 and 2015 present a significant decrease on publications with 2016 showing a small promising increase.

3. Classification of the literature

The amount of the techniques related to social media and their applications in order to spread brand awareness or promote particular products is called Social Media Marketing (SMM). SMM uses mainly the features of social media, such as online communities, social data etc. (Neti, 2011). In the literature, social media marketing is combined with metrics and/or analytics tools, methodologies and techniques. Social media metrics represent the tangible outcome of monitoring, measuring, reporting, calculating content from social media.

Although there is no specific classification system for metrics, researchers can pattern them after: time, reach, relationship, conversion and retention measurements. However, considering that metrics are not yet fully standardized, it depends on the marketer, who sets the marketing goal, to decide the most suitable metric for a certain measurement. Social media produce a vast amount of data, known also as social data, consisting the next phase for an analyst; the social media analysis (SMA). More specifically, SMA consists of gathering and analyzing the data in order to take decisions for businesses. Next, we present eight main definitions for analytics in Table 1.

Furthermore, we classify each article based on five different criteria. More analytically, we subdivide the articles based on methodology, the specific type of analysis, the field of study, the marketing objectives and the social media types/platforms used. As a result, the most common subsectors of each category can be identified, featuring the new upcoming trends on social media marketing. The findings are expected to benefit researchers and marketers by helping them better understand what has been hitherto achieved.

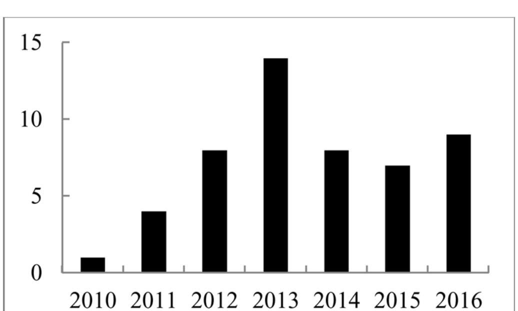


Fig. 1. Year of distribution/number of articles.

Creating a classification constitutes a complex concept to manipulate and conceive, especially in new scientific fields, where literature is still in its early stages. As Bailey (1994) defines, classification is one of the most central and generic of all our conceptual exercises, being the foundation and a necessary process in social science. Typology and taxonomy are two terms that define classification. Typologies are characterized by labels and names. We use the term typology, instead of taxonomy, because our classification system was derived in a deductive manner, without using any cluster analysis or other statistical method, as it occurs with taxonomies. Initially, we did not know which would be our labels, in order to classify the articles. Our selection of articles contained a plethora of labels, which made our mapping process quite complex but challenging. By studying carefully all the articles, we first identified several methodologies, types of analysis, fields of study and marketing objectives. Based on this study, we formed the subsequent Table 2 with the basic labels. This first collection of labels is editable, so future researchers can add, unify or divide the different topics.

Having this classification as a base scheme we study the articles again, this time in order to classify each one in one or more categories. Our scheme lacks of mutual exclusivity, since one article may belong to more than one category. Reviewed articles are classified into five categories and each of them is discussed as follows.

3.1. Methodology of research

Studies follow different approaches related to the methodology used. This depends on the problem's nature and the research field (Noor, 2008). Diverse studies exclusively review the literature. Usually these studies are qualitative and theoretical. We detected 10 articles that perform reviews and/or theoretical research. On the other hand; other studies perform quantitative research using questionnaires. Our study revealed 13 relative articles. The remain articles do not use questionnaires and form the third category of Table 3 with 27 articles.

The generic category of survey-related articles, both questionnaire-based and not, contributes more that 84% of papers. This can be explained by the fact that social media scientists prefer to contribute with primary research articles rather than review-based researches, since the field is quite new and presents a huge research development margin. Though this numeric conclusion can be evidenced by findings, we believe that theoretical approaches are still necessary and form a solid base for conducting primary research.

3.2. Type of analysis

As S3M is a nascent developing field with challenges and opportunities for further research exploration, this Table is designed to assist researchers to obtain the basic knowledge but also to find gaps and limitations, not yet analyzed. The tendencies towards specific research can be revealed also from the next Table. As Gartner (2013) defines, social analytics include sentiment analysis, NLP, text analysis, predictive and content analysis. We enlarge this definition by adding also statistical and behavioral analysis, as possible categories, in our taxonomy. Only one article performs effectuation analysis which is the process for entrepreneurship decision-making (Fischer & Reuber, 2011). These eight categories form the classification for Table 4.

S3M is not yet fully standardized so it is normal that the different categories mix with each other. This is the reason why many papers fit more than one category. Nevertheless, even if classification is not yet fully clarified, we extract the next outcome by observing Table 4. NLP and text analysis, sentiment analysis, content and social media activity analysis are the dominant categories. This observation can be explained by the fact that data contain insights for customers and information for marketers so as to predict useful outcomes.

Table 1
Definitions for Social Media Analytics.

Author(s)	Definitions
Daniel, Hsinchun, Lusch, and Shu-Hsing (2010)	[…] developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application.
Yang, Kiang, Ku, Chiu, and Li (2011)	[…] developing and evaluating informatics tools and frameworks to measure the activities within social media networks from around the web. Data on conversations, engagement, sentiment, influence, and other specific attributes can then be collected, monitored, analyzed, summarized, and visualized.
Mayeh, Scheepers, and Valos (2012)	[…] scanning social media to identify and analyze information about a firm's external environment in order to assimilate and utilize the acquired external intelligence for business purposes.
Grubmüller, Götsch, andKrieger (2013)	[…] social listening and measurements […] based on use generated public content (such as postings, comments, conversations in online forums, etc.)
Sterne and Scott (2010)	SMA is the study of social media metrics that help drive business strategy.
Nielsen (2012)	SMA is the ability to analyze performance of social media initiatives and social data for business intelligence.
Bensen Connie – Dell Company (conniebensen.com)	[…] consist on web analytics, engagement and revenue generated from social media.
Awareness (2012)	[…] an evolving business discipline that aggregates and analyzes online conversation (industry, competitive, prospect, consumer, and customer) and social activity generated by brands across social channels. SMA enable organizations to act on the derived intelligence for business results, improving brand awareness and reputation, marketing and sales effectiveness and customer satisfaction and advocacy.

Table 2
S3 M typology framework for social media metrics and analytics on marketing.

Methodology of research	Literature review and/or Theoretical approach	
	Surveys	Questionnaire-based research Non questionnaire-based research
Type of analysis	Predictive analysis Natural language process (NLP) – Text analysis Effectuation analysis Statistical analysis Sentiment analysis Behavioral analysis Social media activity analysis Content analysis	
Field of study	Banking Education Child welfare and advocacy Tourism industry Stock market Entertainment E-government Food industry Alternative marketing Clothing	
Marketing objectives	Awareness & Branding Engagement eWOM advertising & promotion Predictive marketing research Consumer behavior research Social capital – Value (business, firm equity) – ROI Relationship marketing: CRM & social CRM	
Social media type/ platform	Social networking sites	Facebook, Hi5, Linkedin, Myspace
	Blogs	Blogspot, digg wordpress
	Microblogs	Twitter, twitxr, tweetpeek, plurk
	Content communities – Video sharing sites	Youtube, Flickr, Slideshare
	Forums – discussion	Phpbb. Phorum, skype, messenger, google talk

3.3. Field of study

Our review ends up with 10 specific market fields of study. Only 18 articles clearly focus on specific fields of study, while the rest are generic. We list these fields in Table 5.

As we notice, articles related to tourism industry hold the largest percentage with six articles. In a total of 52 articles, this number represents the 11.5%, but among the 18 that focus on specific categories, the percentage rises in 33.3%. The fact that one third of the articles belong in the tourism industry was highly expected since tourism represents one of the most profitable industries worldwide, contributing to the global economy more than 48 trillion dollars for the same time span of our research, from 2010 to 2016 (Statista.com, 2016).

3.4. Marketing objectives

Business organizations create marketing programs, activities, and campaigns in order to move their current/potential customers to the buyer's journey, designed to align marketing goals and sales activities. Several marketing objectives including specific actions were identified, such as brand awareness, engagement, marketing and especially customer research, behavioral targeting, e-WOM & promotion policy, relationship management & social CRM and social capital value including ROI questions/assessment.

Brand awareness means the exposure of the target audience to brand content and message, while engagement generates further actions taking into consideration the brand content/offers. Marketing and customer research have been identified by marketers as important and common objectives related to the social media use, giving them valuable information regarding customers impressions, sentiment, satisfaction in order to estimate the conversion/purchase potential. Furthermore, the activation of customers' influence based on referrals, advocacy activities and evangelism inspiration for products/services constitutes the marketing objective of a positive eWOM promotion policy. Relationship marketing objectives based on social CRM is the new concern in the marketing world, and with good reason (Hoffman & Fodor, 2010; Pentin, 2011). As social media explode among businesses and customers, monitoring, managing and exploiting the resulting data become essential tasks for almost any marketer. Companies are anxious to meet customers where they are in the social media realm looking for the tools to get involved and gain access. Social CRM software works in conjunction with traditional CRM systems to track customer behavior, as a tool that is part of a social media strategy.

The study of the articles revealed initially 7 marketing objectives supported by social media. Table 6 presents the articles based on each objective they serve.

Engagement, consumer behavior research and relationship marketing represent the most dominant among the other categories with 10, 10 and 8 articles, respectively. All these three categories have the *consumer/customer* as common factor. The consumer-centric marketing was presented as the upcoming trend a few years ago and the current literature and our findings demonstrate that tendency towards that

Table 3
Articles' classification concerning the methodology of research.

Methodology of research		Articles	Percentage of articles/totaln (n/52)
Literature review and/or Theoretical approach		Fan and Gordon (2014); Gayo-Avello et al. (2013); Ghezzi, Gastaldi, Lettieri, Martini, and Corso (2016); Hanna, Rohm, and Crittenden (2011); Malthouse, Haenlein, Skiera, Wege, and Zhang (2013); Neirotti, Raguseo, and Paolucci (2016); Nettleton (2013); Praude and Skulme (2015); Stephen (2016)	19.2%
Survey	Questionnaire-based research	Carim and Warwick (2013); Fischer and Reuber (2011); Godey et al. (2016); Guesalaga (2016); Kim and Ko (2012); Lee, Yen, and Hsiao (2014); Michaelidou, Siamagka, and Christodoulides (2011); Nadeem, Andreini, Salo, and Laukkanen (2015); Paek, Hove, Jung, and Cole (2013); Panagiotopoulos, Shan, Barnett, Regan, and McConnon (2015); Praude and Skulme (2015); Sheth, Sisodia, and Sharma (2000); Tiago and Veríssimo (2014)	25%
	Non questionnaire-based research	Andrew et al. (2012); Asur and Huberman (2010); Braojos-Gomez, Benitez-Amado, and Javier Llorens-Montes (2015); Castronovo and Huang (2012); Chen, Tang, Wu, and Jheng (2014); Geurin and Burch (2016); He, Zha, and Li (2013); Jang, Sim, Lee, and Kwon (2013); Kavanaugh et al. (2012); Kelling, Kelling, and Lennon (2013); Kontopoulos, Berberidis, Dergiades, and Bassiliades (2013); Lau, Li, and Liao (2014); Lieberman (2014); Mostafa (2013); Pehlivan, Sarican, and Berthon (2011); Podobnik (2013); Qiu, Rui, and Whinston (2014); Ribarsky, Xiaoyu Wang, and Dou (2014); Rohm, Milne, and Kaltcheva (2012); Sabate, Berbegal-Mirabent, Cañabate, and Lebherz (2014); Smith, Fischer, and Yongjian (2012); Xiang, Schwartz, Gerdes Jr, and Uysal (2015); Xie et al. (2012); Yadav, de Valck, Hennig-Thurau, Hoffman, and Spann (2013); Yakushev and Mityagin (2014); Yu, Duan, and Cao (2013)	50%

Table 4
Articles' classification concerning the type of analysis.

Type of analysis (primary data collection and/or metric analysis)	Articles	Percentage of articles/total (n/52)
Predictive analysis	Asur and Huberman (2010); Chen et al. (2014); Qiu et al. (2014)	5.8%
Natural Language Process (NLP) − Text analysis	Asur and Huberman (2010); He et al. (2013); Jang et al. (2013); Kontopoulos et al. (2013); Mostafa (2013); Xiang et al. (2015); Yakushev and Mityagin (2014); Yu et al. (2013)	15.3%
Effectuation analysis	Fischer and Reuber (2011)	1.9%
Statistical analysis	He et al. (2013); Podobnik (2013)	3.8%
Sentiment analysis	Chen et al. (2014); Jang et al. (2013); Kontopoulos et al. (2013); Lau et al. (2014); Mostafa (2013); Xiang et al. (2015); Yu et al. (2013)	13.4%
Behavioral analysis	Andrew et al. (2012); Mostafa (2013); Qiu et al. (2014); Xie et al. (2012)	7.7%
Social media activity analysis	Bernabé-Moreno, Tejeda-Lorente, Porcel, Fujita, and Herrera-Viedma (2015); Guesalaga (2016); He et al. (2013); Lieberman (2014); Praude and Skulme (2015); Rohm et al. (2012); Sabate et al. (2014)	13.4%
Content analysis	Bernabé-Moreno et al. (2015); Geurin and Burch (2016); He et al. (2013); Jang et al. (2013); Neirotti et al. (2016); Panagiotopoulos et al. (2015); Ribarsky et al. (2014); Smith et al. (2012); Xiang et al. (2015)	15.4%

Table 5
Articles' classification concerning the field of study.

Field of study	Articles	Percentage of articles/total (n/52)
Banking	Ribarsky et al. (2014)	1.9%
Education	Kelling et al. (2013)	1.9%
Child welfare and advocacy	Paek et al. (2013)	1.9%
Tourism industry	Bernabé-Moreno et al. (2015); Kontopoulos et al. (2013); Mariani, Di Felice, and Mura (2016); Neirotti et al. (2016); Sabate et al. (2014); Xiang et al. (2015)	11.5%
Stock market	Yu et al. (2013)	1.9%
Entertainment (movies, sports)	Asur and Huberman (2010); Geurin and Burch (2016); Podobnik (2013)	5.7%
E-government	Kavanaugh et al. (2012)	1.9%
Food industry	He et al. (2013); Panagiotopoulos et al. (2015)	1.9%
Alternative marketing (viral, email, guerilla etc.)	Castronovo and Huang (2012)	1.9%
Clothing	Nadeem et al. (2015)	1.9%

direction (Osborne & Ballantyne, 2012; Sheth et al., 2000). Of the 47 articles related to some marketing objective, presented in Table 6, the 59.6% regards consumer-centric articles.

3.5. Social media types/platforms and suggested framework

Table 7 represents the articles' distribution for the social media types or the platform used. In order to create this Table, we base our taxonomy on Kaplan and Haenlein (2010), Constantinides and Fountain (2008) and Mangold and Faulds (2009). A difference between these three articles is that the first two use the term *Content communities* for YouTube and the third one, *video sharing sites*.

In a total of 52 articles, 46 of them fit Table 7 with several articles studying multiple social media types or platforms. Articles related to Facebook and Twitter, dominate with 20 and 17 articles, and 38.5% and 32.7% respectively. These results were rather expected, given the fact that 1.86 billion Facebook users and 320 million Twitter users own an active account on these two most visited and diffused SNS and microblog platforms. In global scale, Facebook is used by 54% of global internet users, so it is expected that science will also be of interest for

Table 6
Articles' classification concerning the marketing objectives.

Marketing objectives	Articles	Percentage of articles/total (n/52)
Awareness & Branding	Andrew et al. (2012); Kim and Ko (2012); Lieberman (2014); Mostafa (2013); Rohm et al. (2012); Sabate et al. (2014); Smith et al. (2012)	13.4%
Engagement	Fischer and Reuber (2011); Guesalaga (2016); Malthouse et al. (2013); Mariani et al. (2016); Osborne and Ballantyne (2012); Paek et al. (2013); Panagiotopoulos et al. (2015); Rohm et al. (2012); Sabate et al. (2014)	19.2%
eWOM advertising & promotion	Chen et al. (2014); Stephen (2016)	1.9%
Predictive marketing research	Asur and Huberman (2010); Gayo-Avello et al. (2013); Kim and Ko (2012); Qiu et al. (2014); Yakushev and Mityagin (2014)	9.6%
Consumer behavior research	Bernabé-Moreno et al. (2015); Godey et al. (2016); Jang et al. (2013); Mostafa (2013); Nadeem et al. (2015); Ribarsky et al. (2014); Rohm et al. (2012); Stephen (2016); Xiang et al. (2015); Xie et al. (2012)	19.2%
Social capital – Value (business, firm equity) – ROI	Braojos-Gomez et al. (2015); Fan and Gordon (2014); Godey et al. (2016); He et al. (2013); Lee et al. (2014); Neirotti et al. (2016); Pehlivan et al. (2011); Yu et al. (2013)	15.4%
Relationship marketing: CRM & social CRM	Geurin and Burch (2016); Malthouse et al. (2013); Nadeem et al. (2015); Osborne and Ballantyne (2012); Yadav, Banwari, Parmar, and Maniar (2013)	9.6%

Table 7
Articles' classification concerning the social media types/platforms.

Social media types/platforms	Articles	Percentage of articles/total (n/52)
Social Networking Sites (SNS)	Andrew et al. (2012); Carim and Warwick (2013); Chen et al. (2014); He et al. (2013); Kavanaugh et al. (2012); Kim and Ko (2012); Lee et al. (2014); Lieberman (2014); Mariani et al. (2016); Nadeem et al. (2015); Paek et al. (2013); Podobnik (2013); Ribarsky et al. (2014); Rohm et al. (2012); Sabate et al. (2014); Sheth et al. (2000); Smith et al. (2012); Tiago and Veríssimo (2014); Xie et al. (2012); Yadav, Banwari, Parmar, and Maniar (2013)	38.5%
Blogs	Paek et al. (2013); Yakushev and Mityagin (2014); Yu et al. (2013)	5.8%
Microblogs	Asur and Huberman (2010); Bernabé-Moreno et al. (2015); Carim and Warwick (2013); Fischer and Reuber (2011); He et al. (2013); Kavanaugh et al. (2012); Kelling et al. (2013); Kim and Ko (2012); Kontopoulos et al. (2013); Lieberman (2014); Mostafa (2013); Paek et al. (2013); Ribarsky et al. (2014); Rohm et al. (2012); Sheth et al. (2000); Smith et al. (2012); Yu et al. (2013)	32.7%
Content communities – Video sharing sites	Carim and Warwick (2013); Geurin and Burch (2016); Jang et al. (2013); Kavanaugh et al. (2012); Smith et al. (2012)	9.6%
Forums	Yu et al. (2013)	1.9%

these two platforms.

Summarizing our findings, with respect to the corpus of all articles, we notice a peak on publications in 2013 followed by a decrease the next two years. An important finding is that 2016 represents a small but constant increase in the number of publications, showing an overall increase of interest in social media marketing analysis. Trends show tourism industry, Facebook and Twitter as well as consumer-centric marketing to be the dominant categories, platforms and concepts behind social media marketing strategies. On the other hand though, these trends may bring to the surface gaps in other fields that need attention and research.

4. Future research directions & limitations

Marketing science, together with information technology, has great interest in understanding and analyzing social media and their created data. We present a complete-scale study, aiming to create a typology for social media metrics and analytics related articles, within a continuously incremental and editable typology. The findings contribute to the literature in several ways. The proposed typology is flexible, which means that future literature reviews on a subject can contribute, based on the typology, by adding items and categories.

Regarding specific outcomes, from Table 3 it becomes clear that primary data collection is the most used method on S3 M and the data used for analysis, originate from primary metrics. Another useful outcome is the platform used. On today's social media research, Facebook and Twitter are the dominant platforms and so, the biggest part of the literature is focusing on these two platforms. The above conclusions can help researchers to understand better the tendencies of the diverse field, but also reveal research gaps and lacks on the literature. We believe that the presented paper presents potential for applications in many domains, ranging from marketing to academic or business research. By knowing how to effectively measure the social media value, companies and individuals can produce insights that allow improvement in promoting products and services. Our paper, presents also some limitations. The research was conducted with keywords such as "social media marketing" and not separately for each marketing objective (e.g. branding, engagement, etc.). This fact, limited the number of the articles. Future studies must approach the S3M topic, by searching (and adding as keywords) every field; platform and marketing objective; separately. This type of search will lead to diverse studies; focusing on a specific direction. Our proposed S3M typology framework should trigger future research enabling the incorporation of further criteria.

References

Andrew, L., Mudd, G., Rich, M., & Bruich, S. (2012). The power of like: how brands reach and influence fans through social media marketing. *Journal of Advertising Research*. http://dx.doi.org/10.2501/JAR-52-1-040-052.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *Paper presented at the proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, Vol. 1*.

Awareness, I. (2012). *Actionable social analytics: from social media metrics to business insights.* [Retrieved 14/12/2015, from http://igo2group.com/wp-%C2%AD%E2%80%90content/uploads/2012/10/Actionable-%C2%AD%E2%80%90Social-%C2%AD%E2%80%90Analytics.pdf].

Bailey, K. D. E. (Vol. Ed.), (1994). *Typologies and taxonomies: an introduction to classification: Vol. 102*Sage.

Bernabé-Moreno, J., Tejeda-Lorente, A., Porcel, C., Fujita, H., & Herrera-Viedma, E. (2015). CARESOME. *Know.-Based Syst. 80(C*, 163–179. http://dx.doi.org/10.1016/j.knosys.2014.12.033.

Braojos-Gomez, J., Benitez-Amado, J., & Javier Llorens-Montes, F. (2015). How do small

firms learn to develop a social media competence? *International Journal of Information Management, 35*(4), 443–458. http://dx.doi.org/10.1016/j.ijinfomgt.2015.04.003.
Carim, L., & Warwick, C. (2013). Use of social media for corporate communications by research-funding organisations in the UK. *Public Relations Review, 39*(5), 521–525. http://dx.doi.org/10.1016/j.pubrev.2013.08.006.
Castronovo, C., & Huang, L. (2012). Social media in an alternative marketing communication model. *Journal of Marketing Development & Competitiveness, 6*(1), 117.
Chen, Y.-L., Tang, K., Wu, C.-C., & Jheng, R.-Y. (2014). Predicting the influence of users' posted information for eWOM advertising in social networks. *Electronic Commerce Research and Applications, 13*(6), 431–439. http://dx.doi.org/10.1016/j.elerap.2014.10.001.
Constantinides, E., & Fountain, S. J. (2008). Web 2.0: Conceptual foundations and marketing issues. *Journal of Direct Data Digital Marketing Practice, 9*(3), 231–244.
Fan, W., & Gordon, M. (2014). Unveiling the power of social media analytics. *Communications of ACM*.
Fischer, E., & Reuber, A. R. (2011). Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior? *Journal of Business Venturing, 26*(1), 1–18. http://dx.doi.org/10.1016/j.jbusvent.2010.09.002.
Gartner (2013). *Social analytics.* [from http://www.gartner.com/it-glossary/social-analytics].
Gayo-Avello, P. T. M., Mustafaraj, E., Strohmaier, M., Schoen, H., Gloor, P., Schoen, D., et al. (2013). The power of prediction with social media. *Internet Research, 23*(5), 528–543. http://dx.doi.org/10.1108/IntR-06-2013-0115.
Geurin, A. N., & Burch, L. M. (2016). User-generated branding via social media: an examination of six running brands. *Sport Management Review*. http://dx.doi.org/10.1016/j.smr.2016.09.001.
Ghezzi, A., Gastaldi, L., Lettieri, E., Martini, A., & Corso, M. (2016). A role for startups in unleashing the disruptive power of social media. *International Journal of Information Management, 36*(Part A (6)), 1152–1159. http://dx.doi.org/10.1016/j.ijinfomgt.2016.04.007.
Godey, B., Manthiou, A., Pederzoli, D., Rokka, J., Aiello, G., Donvito, R., et al. (2016). Social media marketing efforts of luxury brands: influence on brand equity and consumer behavior. *Journal of Business Research, 69*(12), 5833–5841. http://dx.doi.org/10.1016/j.jbusres.2016.04.181.
Grubmüller, V., Götsch, K., & Krieger, B. (2013). Social media analytics for future oriented policy making. *European Journal of Futures Research, 1*(1), 1–9. http://dx.doi.org/10.1007/s40309-013-0020-7.
Guesalaga, R. (2016). The use of social media in sales: individual and organizational antecedents, and the role of customer engagement in social media. *Industrial Marketing Management, 54*, 71–79. http://dx.doi.org/10.1016/j.indmarman.2015.12.002.
Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: the power of the social media ecosystem. *Business Horizons, 54*(3), 265–273. http://dx.doi.org/10.1016/j.bushor.2011.01.007.
He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: a case study in the pizza industry. *International Journal of Information Management, 33*(3), 464–472. http://dx.doi.org/10.1016/j.ijinfomgt.2013.01.001.
Hoffman, D., & Fodor, M. (2010). Can you measure the ROI of your social media marketing? *MITSloan Management Review, 52*(1).
Jang, H.-J., Sim, J., Lee, Y., & Kwon, O. (2013). Deep sentiment analysis: mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Systems with Applications, 40*(18), 7492–7503. http://dx.doi.org/10.1016/j.eswa.2013.06.069.
Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*(1), 59–68 [10.1016/j.bushor.2009.09.003].
Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., et al. (2012). Social media use by government: from the routine to the critical. *Government Information Quarterly, 29*(4), 480–491. http://dx.doi.org/10.1016/j.giq.2012.06.002.
Kelling, N. J., Kelling, A. S., & Lennon, J. F. (2013). The tweets that killed a university: A case study investigating the use of traditional and social media in the closure of a state university. *Computers in Human Behavior, 29*(6), 2656–2664. http://dx.doi.org/10.1016/j.chb.2013.06.044.
Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research, 65*(10), 1480–1486. http://dx.doi.org/10.1016/j.jbusres.2011.10.014.
Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications, 40*(10), 4065–4074. http://dx.doi.org/10.1016/j.eswa.2013.01.001.
Lau, R. Y. K., Li, C., & Liao, S. S. Y. (2014). Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems, 65*(0), 80–94. http://dx.doi.org/10.1016/j.dss.2014.05.005.
Lee, M. R., Yen, D. C., & Hsiao, C. Y. (2014). Understanding the perceived community value of Facebook users. *Computers in Human Behavior, 35*(0), 350–358. http://dx.doi.org/10.1016/j.chb.2014.03.018.
Lieberman, M. (2014). Visualizing big data: social network analysis. *Paper presented at the digital research conference*.
Malthouse, E. C., Haenlein, M., Skiera, B., Wege, E., & Zhang, M. (2013). Managing customer relationships in the social media era: introducing the social CRM house. *Journal of Interactive Marketing, 27*(4), 270–280. http://dx.doi.org/10.1016/j.intmar.2013.09.008.
Mangold, W. G., & Faulds, D. J. (2009). Social media: the new hybrid element of the promotion mix. *Business Horizons, 52*(4), 357–365. http://dx.doi.org/10.1016/j.bushor.2009.03.002.
Mariani, M. M., Di Felice, M., & Mura, M. (2016). Facebook as a destination marketing tool: evidence from Italian regional Destination Management Organizations. *Tourism Management, 54*, 321–343. http://dx.doi.org/10.1016/j.tourman.2015.12.008.
Mayeh, M., Scheepers, R., & Valos, M. (2012). Understanding the role of social media monitoring in generating external intelligence. *Paper presented at the australasian conference on information systems (23rd: 2012: geelong, victoria)*.
Michaelidou, N., Siamagka, N. T., & Christodoulides, G. (2011). Usage, barriers and measurement of social media marketing: an exploratory investigation of small and medium B2 B brands. *Industrial Marketing Management, 40*(7), 1153–1159. http://dx.doi.org/10.1016/j.indmarman.2011.09.009.
Mostafa, M. M. (2013). More than words: social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241–4251. http://dx.doi.org/10.1016/j.eswa.2013.01.019.
Nadeem, W., Andreini, D., Salo, J., & Laukkanen, T. (2015). Engaging consumers online through websites and social media: a gender study of Italian Generation Y clothing consumers. *International Journal of Information Management, 35*(4), 432–442. http://dx.doi.org/10.1016/j.ijinfomgt.2015.04.008.
Neirotti, P., Raguseo, E., & Paolucci, E. (2016). Are customers' reviews creating value in the hospitality industry? Exploring the moderating effects of market positioning. *International Journal of Information Management, 36*(6), 1133–1143. http://dx.doi.org/10.1016/j.ijinfomgt.2016.02.010.
Neti, S. (2011). Social media and its role in marketing. *International Journal of Enterprise Computing and Business Systems, 1*(2).
Nettleton, D. F. (2013). Data mining of social networks represented as graphs. *Computer Science Review, 7*(0), 1–34. http://dx.doi.org/10.1016/j.cosrev.2012.12.001.
Nielsen, J. K. (Vol. Ed.), (2012). *Actionable social analytics : from social media metrics to business insights actionable social analytics: Vol. 2015*Business Insights.
Noor, K. B. M. (2008). Case study: A strategic research methodology. *American Journal of Applied Science, 5*(11), 1602–1604.
Osborne, P., & Ballantyne, D. (2012). The paradigmatic pitfalls of customer-centric marketing. *Marketing Theory, 12*(2), 155–172. http://dx.doi.org/10.1177/1470593112441564.
Paek, H.-J., Hove, T., Jung, Y., & Cole, R. T. (2013). Engagement across three social media platforms: An exploratory study of a cause-related PR campaign. *Public Relations Review, 39*(5), 526–533. http://dx.doi.org/10.1016/j.pubrev.2013.09.013.
Panagiotopoulos, P., Shan, L. C., Barnett, J., Regan, & McConnon (2015). A framework of social media engagement: Case studies with food and consumer organisations in the UK and Ireland. *International Journal of Information Management, 35*(4), 394–402. http://dx.doi.org/10.1016/j.ijinfomgt.2015.02.006.
Pehlivan, E., Sarican, F., & Berthon, P. (2011). Mining messages: Exploring consumer response to consumer- vs. firm-generated ads. *Journal of Consumer Behaviour, 10*(6), 313–321. http://dx.doi.org/10.1002/cb.379.
Pentin, R. (2011). *A new framework for measuring social media activity.* Internet Advertising Bureau.
Podobnik, V. (2013). An analysis of facebook social media marketing key performance indicators. *The case of premier league brands. Paper presented at the Telecommunications (ConTEL), 2013 12th International Conference on.* (pp. 26–28).
Praude, V., & Skulme, R. (2015). Social media campaign metrics in latvia. *Procedia − Soc. Behav. Sci. 213*, 628–634. http://dx.doi.org/10.1016/j.sbspro.2015.11.462.
Qiu, L., Rui, H., & Whinston, A. B. (2014). Effects of social networks on prediction markets: Examination in a controlled experiment. *Journal of Management Information Systems, 30*(4), 235–268. http://dx.doi.org/10.2753/mis0742-1222300409.
Ribarsky, W., Xiaoyu Wang, D., & Dou, W. (2014). Social media analytics for competitive advantage. *Computers & Graphics, 38*(0), 328–331. http://dx.doi.org/10.1016/j.cag.2013.11.003.
Rohm, A., Milne, G. R., & Kaltcheva, V. (2012). The role of online social media in brand-consumer engagement. *Paper presented at the Direct/Interactive marketing research summit proceedings*.
Sabate, F., Berbegal-Mirabent, J., Cañabate, A., & Lebherz, P. R. (2014). Factors influencing popularity of branded content in Facebook fan pages. *European Management Journal, 32*(6), 1001–1011. http://dx.doi.org/10.1016/j.emj.2014.05.001.
Sheth, J. N., Sisodia, R. S., & Sharma, A. (2000). The antecedents and consequences of customer-centric marketing. *Journal of the Academy of Marketing Science, 28*(1), 55–66. http://dx.doi.org/10.1177/0092070300281006.
Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across youTube, facebook, and twitter? *Journal of Interactive Marketing, 26*(2), 102–113. http://dx.doi.org/10.1016/j.intmar.2012.01.002.
Statista.com (2016). *Direct and total contribution of travel and tourism to the global economy from 2006 to 2015 (in trillion U.S. dollars).* [Retrieved 27.2.2017, 2016, from Direct and total contribution of travel and tourism to the global economy from 2006 to 2015 (in trillion U.S. dollars)].
Stephen, A. T. (2016). The role of digital and social media marketing in consumer behavior. *Current Opinion in Psychology, 10*, 17–21. http://dx.doi.org/10.1016/j.copsyc.2015.10.016.
Sterne, J., & Scott, M. D. (2010). *Social media metrics: How to measure and optimize your marketing investment.* Wiley.
Tiago, M. T. P. M. B., & Veríssimo, J. M. C. (2014). Digital marketing and social media: Why bother? *Business Horizons, 57*(6), 703–708. http://dx.doi.org/10.1016/j.bushor.2014.07.002.
Xiang, Z., Schwartz, Z., Gerdes, J. H., Jr, & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*(0), 120–130. http://dx.doi.org/10.1016/j.ijhm.2014.10.013.
Xie, Y., Cheng, Y., Honbo, D., Zhang, K., Agrawal, A., & Choudhary, A. (2012). Probabilistic macro behavioral targeting. *Paper presented at the Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*.
Yadav, M. S., de Valck, K., Hennig-Thurau, T., Hoffman, D. L., & Spann, M. (2013a). Social commerce: A contingency framework for assessing marketing potential.

Journal of Interactive Marketing, 27(4), 311–323. http://dx.doi.org/10.1016/j.intmar.2013.09.001.

Yadav, P., Banwari, G., Parmar, C., & Maniar, R. (2013b). Internet addiction and its correlates among high school students: A preliminary study from Ahmedabad, India. *Asian Journal of Psychiatry, 6*(6), 500–505. http://dx.doi.org/10.1016/j.ajp.2013.06.004.

Yakushev, A., & Mityagin, S. (2014). Social networks mining for analysis and modeling drugs usage. *Procedia Computer Science, 29*(0), 2462–2471. http://dx.doi.org/10.1016/j.procs.2014.05.230.

Yang, M., Kiang, M., Ku, Y., Chiu, C., & Li, Y. (2011). Social media analytics for radical opinion mining in hate group web forums. *Journal of Homeland Security and Emergency Management, Vol. 8*.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919–926. http://dx.doi.org/10.1016/j.dss.2012.12.028.

Active Viral Marketing: Incorporating Continuous Active Seeding Efforts into the Diffusion Model

Alon Sela[a,b], Dmitri Goldenberg[a], Irad Ben-Gal[a], Erez Shmueli[a,*]

[a]*Department of Industrial Engineering, Tel-Aviv University, Tel-Aviv, Israel*
[b]*Department of Industrial Engineering, Ariel University, Ariel, Israel*

Abstract

Existing viral-marketing network models commonly assume a preliminary phase in which a marketer actively infects a subset of social network's users, represented by nodes, followed by a passive viral process, in which nodes infect other nodes without external intervention. However, in real-world commercial scenarios, substantial efforts are often invested by companies to promote their products, suggesting that the adoption of products is rarely the consequence of a viral spread alone.

Under this observation, this paper proposes a new diffusion model, named Active Viral Marketing, which better fits real-world marketing scenarios, where adoption of products relies on continuous active promotion efforts by the marketer. In the proposed model, the success of a marketing attempt to infect a potential customer (uninfected node), depends on the number of adopting friends (infected neighbors) of this user, assuming a user is more likely to adopt a product if more of his/her friends have already adopted it, while taking into account that social influence diminishes over time due to a memory-loss effect.

The paper further proposes a set of heuristics to schedule the marketing attempts. The main idea behind these heuristics is to utilize the information on the dynamic adoption-states of neighbor nodes, in addition to the static social network topology, when choosing the next node to seed. An extensive experimentation demonstrates how the proposed seeding heuristics improve

*Corresponding author

Email addresses: alonse2012@gmail.com (Alon Sela), dmitrig@mail.tau.ac.il (Dmitri Goldenberg), bengal@tau.ac.il (Irad Ben-Gal), shmueli@tau.ac.il (Erez Shmueli)

the adoption rate of products by 30%-75% in comparison to existing state-of-the-art methods that mainly rely on the network topology.

Keywords: Influence Maximization; Information Diffusion; Viral Marketing; Scheduled Seeding

1. Introduction

Online social networks offer a powerful tool for information sharing with friends, family and colleagues. In this aspect, they enable individuals to spread their messages passively through a viral process that might resemble the spread of a virus. Clearly, this property of online social networks also has a financial implication, since it can be utilized by companies and individuals that seek to advertise their products (we are using the terms products and services, interchangeably) to reach a large number of potential customers.

The importance of social influence in information spread processes was demonstrated in many studies (e.g., Asch (1951); Centola & Macy (2007)). One of the traditional models for describing diffusion of information in networks is the Linear Threshold model (Granovetter, 1978). In this model, a message spreads from one node to another if a fixed number of neighbors of the latter have already adopted it. The spreader of the message, who is interested to reach a large number of adopting nodes, has to wisely select a subset of network nodes, and actively infect (seed) them. Then, the assumption is that a passive viral process begins, in which nodes infect other nodes without any external intervention. Such a passive viral process can happen for example, if a Facebook user posts an exciting message or photo on his wall, which is then repeatedly shared by other Facebook users.

As shown by both analytical and simulative studies (Barthélemy et al., 2004; Khelil et al., 2002; Vespignani, 2012; Zhou et al., 2007), messages that propagate according to models similar to the Linear Threshold model, are expected to propagate into a substantial portion of the network. However, in recent years, several works (e.g., Leskovec et al. (2009, 2007b); Leskovec & Horvitz (2008); Goel et al. (2012)) have shown, based on real information cascades datasets, that the frequency of large information cascades in networks, are significantly lower than what was previously believed. In fact, it was shown that the vast majority of messages never spread beyond a few nodes.

Since large information cascades are rare in reality, it is unlikely that a product or service will be spread only through a passive viral process. Indeed, in most real-world marketing scenarios, substantial additional efforts are invested in order to promote products. Companies cannot simply post Facebook messages on their products and expect them to spread passively, and therefore so many sales and marketing personnel are hired to actively promote these commercial products and services.

In this work, we propose a new information diffusion model, named Active Viral Marketing (AVM), which better reflects the need of commercial companies to invest continuous marketing efforts to promote their products. More specifically, nodes in our model cannot get infected by themselves through a passive viral process. Instead, they can get infected only through an active seeding attempt made by the spreader. The importance of social influence comes into play where the success of a seeding attempt depends both on the number of infected neighbors the node has and on the time frame in which they got infected.

As a motivating example, consider a tourism company that aims to promote a summer vacation through social networks advertising. The company can pop-up the advertisement to several social network's users that fit in age and social class and are considered to be influential (commonly estimated based on the network topology). These advertisements have a defined cost, which is paid to the social network company (and sometimes to the influencers as well). It is likely that a user that already clicked on the ad (and potentially booked a vacation), will discuss it with his social network's friends. However, in most cases it is unlikely that following such a discussion, these friends will initiate a call to the tourism company (as in the passive viral spread) in order to book a vacation. Rather it is more likely that the discussion with friends that already booked a vacation will have some positive impact on them. Now, consider a user that had several such discussions with his friends. If the company chooses to pop-up an ad to this user now, the likelihood of him clicking on the ad will be probably higher than in the case without previous discussions. Moreover, if the ad is presented to the user, long after the discussions with his friends, it is less likely that he will click on the ad, since the impact of these interactions weakens over time (for simplicity of exposition, we will term this phenomenon as "forgetting effect" or as "memory loss"). Therefore, in order for the tourism company to efficiently use its marketing budget (e.g., to maximize the click-through rate), it needs to schedule its advertisements in a way that balances the number of accumulated friends' clicks and memory loss (both grow over time and have an inverse effect).

Following the suggested AVM model and the observation above, this work develops a set of Scheduled Seeding Heuristics (SSH). The main idea behind SSH is to utilize the information on the dynamic states of nodes, in addition to the static network topology (that is commonly used by existing seeding heuristics), when choosing the next node to seed. This added information allows SSH to utilize better the social effect, by balancing between the number of infected neighbors of a node and its memory loss.

In order to evaluate the SSH heuristics, we conducted an extensive set of experiments, to compare them to other state-of-the-art seeding heuristics that rely on selecting central nodes (based on the network topology) prior to the seeding stage. The results of our experiments show that the SSH heuristics obtain an average adoption rate which is 30%-75% higher than the other benchmark heuristics, and that the superiority of SSH is consistent over a wide range of parameters' values selection.

The contribution of this work can be summarized along two axes:

- We propose a new diffusion model which, to our belief, better fits real-world scenarios of products adoption, where the spread of products relies on continuous active efforts of the sales or marketing departments.
- We demonstrate the importance and the high potential of a scheduled seeding heuristic, for the spread of trendy products, under a wide range of settings, and also point out the cases where such a heuristic is less effective.

The rest of this paper is structured as follows. Section 2 reviews the existing literature and provides the necessary background on information diffusion in networks. Sections 3 and 4 describe the proposed Active Viral Marketing model and Scheduled Seeding Heuristics, respectively. Section 5 details our evaluation methodology and the obtained results. Section 6 summarizes the paper, and presents directions for future research.

2. Background and Related Work

In this section, we provide the relevant background to the fields of contagion models and viral marketing. We start by presenting the basic theoretical models of viral diseases, followed by two well-known models, which capture the important aspects of viral marketing. These theoretical models are then inspected through the lens of real-world data evidences.

2.1. Contagion Models

Mathematical contagion models of diseases were historically developed by Epidemiology researchers as a tool to study the mechanisms by which diseases spread, to predict the future course of an outbreak and to evaluate strategies to control an epidemic (Anderson et al., 1992). Due to their success in the field of disease modeling, such models implied their wide usage in other fields as well, such as information diffusion and product adoption.

Existing contagion models can be broadly classified into two categories: (1) compartmental models and (2) individual-based models.

Compartmental models assume a fully interconnected population, in which the interactions and infections can occur between any pair of available individuals. This implies a homogeneous population in terms of their connectivity and chances of interaction. These models allow to observe different phenomena at the compartment level, such as the size of the compartment and the infection pace at different time periods of the contagion process. One of the most well-studied compartmental contagion models is the SIR model (Anderson et al., 1992). This model splits the population individuals into three compartments: S - susceptible, I - Infected and R - Recovered. The transitions between the states in this model are trivial - susceptible individuals have a probability β to become infected as a result of an interaction with infected individuals. Similarly, infected individuals recover (and therefore reassigned into the recovered compartment) with a constant pace γ.

Individual-based models assume the existence of a network structure that describe the potential interactions (network edges) between individuals (network nodes). In contrast to compartmental models, individuals cannot become infected from any member of the infected compartment, but only from their network neighbors.

One of the fundamental individual-based models, commonly used to describe information diffusion in social networks is the Linear Threshold model (Granovetter, 1978; Kempe et al., 2003). The model assumes that the behavior of individuals greatly depends on the number of their network neighbors that are already engaged in that behavior. More formally, we denote the binary state of a node v (1 if active and 0 otherwise) at time t as $X_v(t)$ and the set of neighbors of node v as $N(v)$. A node v is influenced by each neighbor $w \in N(v)$ according to their edge weights $b_{v,w}$ which are set such that $\sum_{w \in N(v)} b_{v,w} = 1$. Each node v is assigned a threshold $\theta_v \in [0, 1]$, representing the fraction of v's neighbors that are required to be active in order for v to become active in the next time step. If the accumulated effect (sum of weights of active neighbors) on time step t on v is at least θ_v, v will become active at the next time step $t + 1$ and therefore will also begin to influence its own neighbors.

Another well-studied individual-based information diffusion model is the Independent Cascade model (Goldenberg et al., 2001a,b). In this model, a node v that was activated at time step t has a single chance to activate each of its currently inactive neighbors $w \in N(v)$. At the next time step, $t + 1$, v will not have any further influence on its neighbors. Similarly, if w becomes activated at time step $t + 1$, it will have one single chance to activate its inactive neighbors in time step $t + 2$.

A particularly interesting individual-based model, Bass-SIR, was recently suggested by Fibich (2016). This model proposes a new contagion process which combines properties of SIR and Bass (Mahajan et al., 1991) models, and applies them at the micro-level by utilizing a network structure. More specifically, as in the basic Bass model, if a node v did not adopt the product by time step t, it has a positive probability to adopt the product in the nearest future $(t, t + \Delta t)$:

$$P(v \text{ adopts in } (t, t + \Delta t)) = (p + q\frac{I_v(t)}{k_v})\Delta t + o(\Delta t)$$

Where p and q are bass coefficients of innovation and imitation accordingly, $I_v(t)$ is the number of infective neighbors of v at time step t, k_v is a normalization factor (usually $k_v = |N(v)|$) and $\Delta t \to 0$. Unlike the basic Bass model, Bass-SIR does not assume that an infected node will stay infective forever, and therefore the probability of an infective node to become recovered is:

$$P(v \text{ recovers in } (t, t + \Delta t)) = r\Delta t + o(\Delta t)$$

Where $\Delta t \to 0$, and r is the recovery pace.

The Linear Threshold and Independent Cascade models served as a basic setup to a wide range of works, and over the years many extensions were suggested to fit these models to special cases. In their seminal work, Kempe et al. (2003) proposed two models which aimed at generalizing many of the extensions into a unified framework. The introduction of these two general models served several goals. First, they present a unified framework for any arbitrary activation function that is consistent with the monotonicity condition. Second, they prove that these two models are equivalent, and provide a method to covert between them. Third, when limiting the discussion to sub-modular activation functions, Kempe et al. provide an approximation to the Influence Maximization problem, covered later in section 2.2.

2.2. Influence Maximization

An important field in the study of information diffusion through social networks is the identification of influential nodes with the goal of maximizing the adoption of products or ideas in the network. More formally, given a model of information diffusion (e.g., Linear Threshold, Independent Cascade, etc.) over a network G, the influence maximization problem deals with selecting a subset of the network nodes, whose intentional activation (often referred to as seeding) will ignite a viral contagion process that will impact a significantly large set of nodes. Often these models aim at optimizing a given target function related to the network adoption. The target function can have several forms, such as maximizing the number of adopters in a certain time period or budget (number of seeding actions), or minimizing the number of seeding actions required to reach a certain number of adopters.

For example, modern marketing efforts use social networks for market analysis and for defining promotion strategies. Unlike classical mass-marketing methods that address a wide market segment, social networks' promotion is often characterized by micro-segmentation, attempting to utilize detailed information about each of the involved individuals (Goldfarb & Tucker, 2011). The main motivation behind such an approach, is that influencing the opinion of only a few individuals may shape the opinion of the majority, by following a viral contagion process (Katz & Lazarsfeld, 1955).

The task of identifying influential nodes is still widely investigated, but the identification of influential nodes is not always easy. In many cases, nodes are referred to as "influential" when past evidence show that their involvement in the contagion process contributes significantly to the spread. Nonetheless, such detailed information is often absent, and most of the data available to the marketers is the topological structure of the social network and past adoption history.

2.2.1. Initial Seeding Strategies for Influence Maximization

Identifying influential nodes, given only the network structure, can be addressed via graph-based metrics, such as the centrality measures (Borgatti, 2005).

One way to measure a node's centrality is by counting the number of its connections (known as the node degree). While calculating the degree of a node is a relatively trivial task, such an approach is limited since it takes into account only the first-order effect, without considering higher-order effects. Other frequently used centrality measures that take into account high-order effects include the PageRank (Page et al., 1999), the Betweenness centrality (Brandes, 2001) and the Eigenvector centrality (Bonacich, 2007). Each of these measures has its own attributes and represents a different type of importance that characterizes a node. For a good source on centrality measures, the reader is referred to (Borgatti, 2005) and (Newman).

With respect to influence maximization, several works investigated the efficiency of seeding central nodes. The work by Hinz et al. (2011), for example, investigated four seeding strategies: Hubs (Degree/EigenVector Centrality), Bridges (Betweenness Centrality), Fringes (Edge Nodes) and Random. The authors conducted three experimental studies of adoption using a small controlled network; a real social network of selected students; and a large-scale cellular network. The study found that targeting Hubs is the most effective strategy in terms of influence maximization, with the Bridges strategy right afterwards, both with a big gap above the Random strategy (150-200%) and a huge gap above the Fringes strategy. Similar results were obtained by Banerjee et al. (2013), where the authors investigated empirically the spread of financial loan systems within a social network of Indian villagers. The authors found that villagers with high Eigenvector centrality scores are more likely to influence others in their surroundings, in comparison to the other measures of centrality.

The performance of seeding strategies depends not only on the properties of the network topology and its nodes, but also on the information diffusion dynamics themselves. For example, Kempe et al. (2003) study the influence maximization problem under the linear threshold and independent cascade settings and their generalizations. The authors prove that finding the optimal solution to the problem is NP-hard in both settings and present a greedy algorithm which obtains a $(1 - 1/e)$ approximation of the optimal solution. While the greedy algorithm ensures a reasonably good result in terms of coverage, it is still very expensive in terms of runtime when executed on large-scale datasets.

The complexity of the problem and the non-scalability of the greedy approximation algorithm opened the chase after high performing and scalable seed selection heuristics. While many such heuristics were suggested in the literature, we focus on two well-studied groups of such heuristics.

One notable group of such heuristics are the *CELF* (Leskovec et al., 2007a) and *CELF++* (Goyal et al., 2011) algorithms, which are based on a "lazy-forward" optimization scheme for selecting the seeds. Their underlying idea is based on bounding the marginal contribution of a node in a future iteration, with its marginal contribution in a previous iteration due to monotonicity and sub-modularity properties of the influence maximization problem. These heuristics provide an efficient variation of the greedy approximation algorithm by improving the order of evaluating nodes to be added to the "seed set". Empirical evaluation showed that the proposed heuristics outperform (in terms of influence maximization) and run faster than the greedy algorithm, while still guaranteeing a constant factor approximation of the optimal solution.

Another notable group of heuristics was suggested by Chen at al. (Chen et al., 2009, 2010a; Jung et al., 2012; Chen et al., 2010b). Chen et al. (2009) presented an improved greedy algorithm for seeding outcome evaluation by reducing the search space per each evaluation, and showed a 700-times faster performance on the independent cascade model. Chen et al. (2010a) suggested the Maximum Influence Path (PMIA) algorithm. Using this method under the independent cascade model, the authors suggested to locate the nodes whose seeding will result in a long chain of cascades with the highest probability. Jung et al. (2012) proposed the Influence Rank Influence Estimation (IRIE) algorithm, which performs an estimation of the influence function for any given seed set, using precomputed influence estimated values for iterative seed set ranking. Empirical simulations have shown that the IRIE heuristic performance is similar to that of the Greedy, PMIA and Pagerank influence heuristics, while its memory consumption provides a significant improvement over that of the other heuristics.

While a large number of works in this field focused on the problem of maximizing influence with a given seeding budget, Long & Wong (2011) investigated the problem of minimizing the number of seeding actions to obtain a certain number of influenced nodes. The authors proved that the problem is NP-hard, and developed a greedy heuristic that provides error guarantees. They also studied the "Full-Coverage" setting, where the goal is to infect the entire network, and designed efficient algorithms for this purpose.

With the same spirit, Goyal et al. (2013) identified three orthogonal dimensions in the influence maximization problem: (1) the number of seed nodes activated at the beginning, (2) the expected number of activated nodes at the end of the propagation, and (3) the time taken for the propagation, claiming that it is possible to constrain either one or two of these dimensions and try to optimize the third. The authors then studied two of these variations and suggested approximated algorithms to solve them efficiently.

2.2.2. Adaptive Seeding Strategies for Influence Maximization

The majority of existing works that dealt with the influence maximization problem, focused on selecting a subset of network nodes, that if seeded simultaneously at the beginning of the process, would maximize the adoption rate at the end of the process. Recently, numerous works presented a new adaptive approach, which spreads the seeding actions over time, and therefore allows to reassess the contribution of the seeds' selection in each time step, in order to improve the overall adoption rate.

For example, Seeman & Singer (2013) present a two-stage framework for influence maximization. The underlying assumption of this model is that besides of the "non-active" (susceptible) and "active" (infective) states there is an intermediate state referred to as "available": a node v is considered available for seeding only if one of its neighbors $w \in N(v)$ is active. Given an initial set of available nodes $X \subseteq V$, the goal of the first stage is to select a seeding set $S \subseteq X$ in order to extend the set of available nodes, so that the seeding actions in the second stage will maximize the expected influence. The idea behind it relies on the known fact that selecting a neighbor of a random node v is likely to have a higher degree than v itself and thus one would like to include those higher-degree nodes in the set of available nodes for seeding.

In another study, Tong et al. (2017) suggest an adaptive seeding strategy for a variant of the Independent Cascade model. In this variant, referred to as "Dynamic Independent Cascade" model, the authors assume that the activation of a node v by seeding occurs with a probability p_v. Therefore, in contrast to the models surveyed above, a seeding action may fail, keeping the node in a non-active state. Under this setting, the authors suggest an adaptive seeding approach, in which the selection of nodes to be seeded at each time step, is performed while taking into account the realization of the previous seeding attempts.

Jankowski et al. (2017a,b) suggest an adaptive seeding approach to the influence maximization problem under the Independent Cascade model. The authors show that, regardless of the chosen strategy for selecting influential nodes, spreading the seeding actions along different time-steps of the diffusion process can improve the overall adoption rate. Moreover, they present an inherent trade-off between the obtained adoption rate and the duration of the diffusion process.

Chierichetti et al. (2014) introduce a different diffusion model in which there are two competing ideas, each aiming at maximizing its spread over a social network. More specifically, consider a marketer which addresses each one of the individuals in the network sequentially (the marketer has the ability to determine this sequence) and offers them a cause. The cause can either be accepted (Y) or denied (N) by each of the individuals, according to the following rule: the individual v accepts the offer if $|m_Y| - |m_N| \geq c$, deny it if $|m_N| - |m_Y| \geq c$ and chooses randomly between Y and N otherwise. m_Y and m_N represent the size of the group of v's neighbors who already decided to accept or deny the cause (Y or N), and c is a positive integer that serves as a decision threshold. The goal of the marketer in this setting is to determine the best order to address the individuals in order to maximize the amount of Y decisions. The authors also provide an efficient greedy algorithm that ensures the best achievable solution to the problem.

Lin et al. (2014) suggest the "Push-Driven Cascade" model in which the probability that a node will become active after a seeding action is determined by the activation state of its neighbors. More specifically, the probability of an individual v to become activated is:

$$p_v(t) = d_v + \sum_{w \in N(v)} b_{v,w} * X_w(t-1)$$

Where $X_w(t-1)$ is the binary state of node w (1 if active and 0 otherwise) at time $t-1$, the node v is influenced by each active neighbor $w \in N(v)$ according to their edge weights $b_{v,w}$ and d_v is v's own bias towards adoption. The role of the marketer in this setting is to choose a single node to seed at each time step in order to maximize the overall adoption in the network.

It is important to emphasize that in the two latter models, each node has an accumulated influence in favor of the product, but only the seeding act itself is considered to be the trigger for activation, where the viral spread serves only as a positive effect on the activation probability. This is in contradiction to classical diffusion models where nodes could become active as a result of a viral infection without any external intervening operation.

2.3. Information Diffusion in Real World Settings

As seen in the previous section, the dynamics of information diffusion in Social Networks were widely studied and many mathematical models which aim at describing these dynamics were suggested. In recent years, due to the increased availability of data, and the emergence of tools to store and process data at large-scale, a growing body of works have started to analyze the dynamics of information diffusion in real-world scenarios, and obtain better understanding of where existing models succeed and fail in describing these dynamics.

One of the principles behind many of these models is that of accumulated social effect. Already in 1951, the social psychologist Asch presented an experiment, in which he showed that the probability of a subject to change his opinion is proportional to the number of peers who are convincing him to do so (Asch, 1951). Granovetter (1978) in turn, presented a threshold behavior, in which an accumulated social effect is turned into an activation by reaching a personal threshold of the individual. Hence, since the threshold values are distributed randomly, the probability of an activation is proportional to the number of social influencers, similarly to Asch's findings. Later on, Centola & Macy (2007) had performed a large-scale empirical study of online social networks. He found that in contradiction to "Simple Contagion" in which a single interaction with an infected individual may lead to activation (e.g., like in the spread of infectious diseases), the activation of an individual often requires reinforcement from multiple infected sources, a phenomenon named by the author as "Complex Contagion".

A recent work by Goyal et al. (2010) studied the time effect of propagation of social influence in networks. Consequently, the authors suggested an extension to the General Threshold model by adding a diminishing time-dependency factor. More specifically, they considered three types of time-dependent models which reflect a lower ability of a node to spread the adopted idea as time passes: (1) A Static Model the influence of an infective node does not diminish over time; (2) A Discrete Model each activated node has a period of time in which it is infective. After that period, the node stops from being infective; and (3) Continuous Model the influence of an infective node

v on a neighbor node w diminishes over time with an exponential rate. The authors found that the best fit to the data was obtained by the continuous (exponential decay) model. One explanation that was given to this diminishing influence effect in the scientific literature is the limited attention effect. According to this effect, a person which is exposed to multiple ideas during a single time period, is able to concentrate only on a few of them resulting in a forgetting effect (Weng et al., 2012). These findings, strengthen the usage of the recovery effect in several of the models mentioned above, such as SIR and Independent Cascade.

In another paper by Leskovec et al. (2007b), the authors investigate the cascading behavior of online information diffusion, by analyzing 45,000 blogs and about 2.2 million blog posts. The authors identified several cascade shapes that rule the majority of cascades, pointing out two specific shapes: star-shaped, reflecting the spread of information in different directions, and chain-shaped, presenting a chained sequence of information flow. Further investigating the degree-distribution of the cascades, they found that in-degree and out-degree distribution of bag-of-cascades follow power-law exponents of -2.2 and -1.92 respectively. Finally, by examining the distribution of cascade sizes for each shape of cascade, they found that all cascades follow a heavy-tailed distribution, and the probability of observing a cascade of n nodes follows a Zipf distribution. These findings emphasize that in real-world scenarios, highly viral information cascades rarely exist.

Another support for the above findings can be found in (Goel et al., 2012), where the authors analyze information cascades in seven different online domains. The authors observed that the vast majority of cascades are small, and that they usually terminate within one circle of neighbors of the initial adopting node.

3. The Proposed Active Viral Marketing Model

In this section, we propose a novel information diffusion model, named the Active Viral Marketing model, which better reflects the need of commercial companies to invest continuous marketing efforts to promote their products or services. According to the proposed model, at any given time-step t, a node v can only be at one of the following $X_v(t)$ states:

- $X_v(t) = 0$: Non-Infected
- $X_v(t) = 1$: Infected and Infectious

- $X_v(t) = 2$: Infected but not Infectious
- $X_v(t) = 3$: Seeding Failed

The possible transitions of a node v between these states are described in Figure 1:

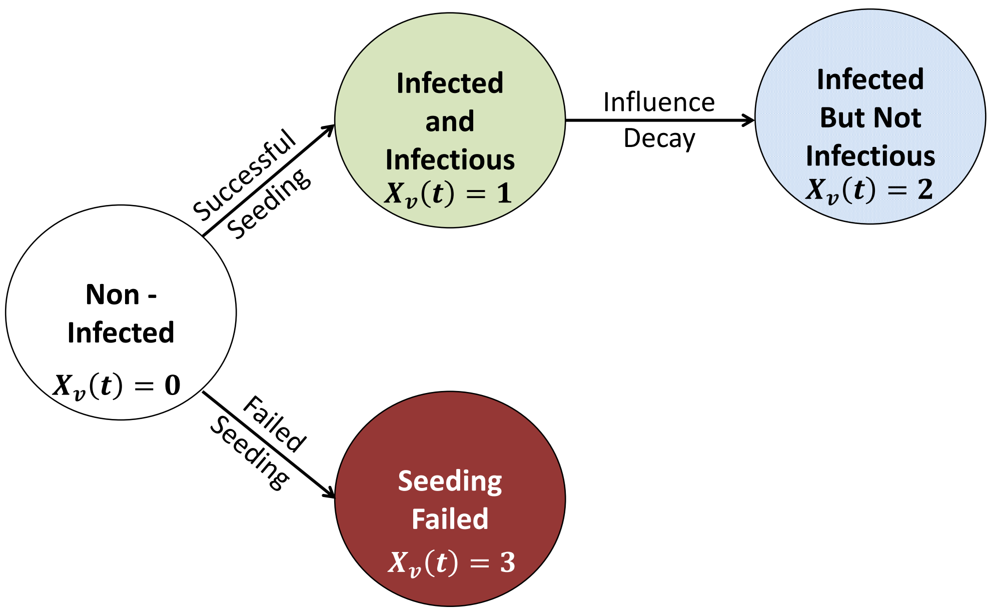


Figure 1: Infection states of nodes in the AVM model.

More specifically, if the spreader attempts to seed a non-infected node v at time-step t, the attempt may succeed with a probability $P_v(t)$. If the seeding attempt succeeds, then the node's state changes from $X_v(t) = 0$ to $X_v(t) = 1$. The probability of a successful seeding attempt is affected by v's individual preferences and the activation rate of v's neighbors (described in more details below).

If the seeding attempt fails, subsequent attempts to seed v are not allowed (since in a typical marketing scenario, subsequent seeding attempts may only annoy the potential customer and may lead to a negative attitude towards the spreader), and v is transitioned into a "Seeding Failed state ($X_v(t) = 3$). On the other hand, if the seeding attempt succeeds, v is transitioned into a “Infected and Infectious” state, and will influence its neighbors only for the next t_{inf} periods. After t_{inf} periods have ended, v's state changes to “Infected but not Infectious” ($X_v(t) = 2$).

The probability that an attempt to seed a node v at time-step t will succeed is given in Eq. 1:

$$P_v(t) = P_v^{ind} + P_v^{soc} \cdot min(1, \frac{|N_v^1(t)|}{\theta_v}) \tag{1}$$

This probability is composed of two factors: (1) the individual preferences of v, denoted by P_v^{ind} and (2) the social influence exerted on v by its infectious neighbors at time-step t, denoted by $P_v^{soc} \cdot min(1, \frac{|N_v^1(t)|}{\theta_v})$.

The social factor is calculated as the product of P_v^{soc} and $min(1, \frac{|N_v^1(t)|}{\theta_v})$. The maximal social effect that can be achieved is represented by P_v^{soc}, (note that $P_v^{ind} + P_v^{soc} \leq 1$). $min(1, \frac{|N_v^1(t)|}{\theta_v})$ represents the relative social effect, which increases proportionally with $|N_v^1(t)|$, denoting the number of infectious (state 1) neighbors of v, up to a certain level determined by the threshold θ_v. The min function assures that even if the number of active neighbors exceeds the threshold θ_v, the probability function would not exceed the value of 1, and therefore, the total social effect would not exceed P_v^{soc}.

The formulation of the social factor described above was inspired by the empirical results of Asch's conformity experiments (Asch, 1951). In his experiments, Asch inspected how the size of a group influences the probability of conforming to the opinion of the majority. He observed that as the size of the group grows, the conforming probability grows almost linearly until reaching a certain size, and after reaching that size, the probability doesn't grow further. We model these two properties by using the threshold θ_v and the maximum probability P_v^{soc}. A comparison of Asch's original findings and our simplified model (for the case of $p_v^{soc} = 0.6$, $p_v^{ind} = 0.1$ and $\theta_v = 4$) are depicted in Figure 2.

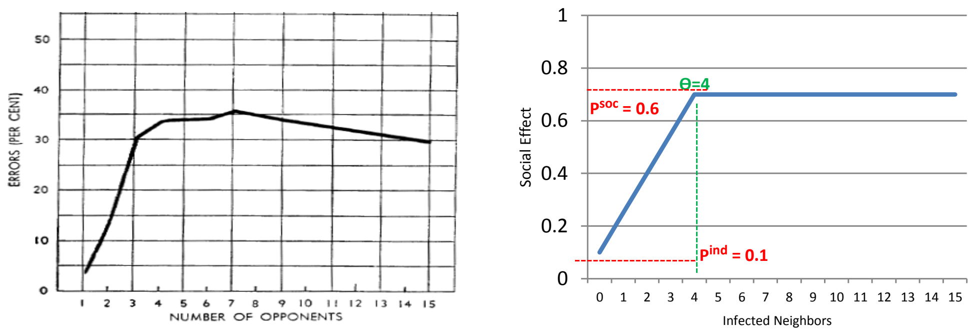


Figure 2: Social effect in Asch's conformity experiment (left) and its representation in the AVM model (right).

Given the Active Viral Marketing diffusion model and a seeding budget of size B, the goal is to find an ordered set of B nodes, denoted by $S = (v_1, v_2, .., v_B)$, such that seeding the node v_1 at time-step $t = 1$, the node v_2 at time-step $t = 2$, ..., the node v_B at time-step $t = B$, would maximize the total number of successful seeding attempts.

4. The Scheduled Seeding Heuristics

The influence maximization problem that was defined above for the Active Viral Marketing diffusion model is NP-hard and is not sub-modular (see Appendix A). Accordingly, in this section, we propose a set of seeding heuristics, named Scheduled Seeding Heuristics (SSH), that recommend which node to seed at each time-step. Similar to existing seeding heuristics, our heuristics utilize the static network topology when choosing the nodes to be seeded. However, in contrast to existing heuristics, our heuristics also take into account the information on the dynamic states of nodes at each time-step.

More specifically, at each time-step, our heuristics assign a utility score for each one of the non-infected (state 0) network nodes, with the idea that seeding a node with a higher utility score is worthier. The utility score is based on the expected value for each potentially seeded node, and is calculated as the probability of a successful seeding of the node itself, multiplied by the value of such an event.

Given the vector of states of all network nodes at time-step t, denoted by $\vec{X}(t)$, the probability of a successful seeding of v at time-step t is denoted as:

$$P(\vec{X}^v(t+1))$$

Where $\vec{X}^v(t+1)$ is identical to $\vec{X}(t)$ with the additional assumption that node v changed its state to $X_v(t) = 1$ at time-step $t+1$.

The value of a successful seeding event of node v can be seen as the influence of v on future seeding attempts of its non-infected neighbors, formulated as:

$$\sum_{w \in N_v^0(t+1)} U(w, t+1, \vec{X}^v(t+1))$$

Where $w \in N_v^0(t+1)$ is a non-infected neighbor of node v at time-step $t+1$, and $U(w, t+1, \vec{X}^v(t))$ is the utility score of seeding w at time-step $t+1$, given that v was already seeded successfully at time-step t.

Finally, the utility score of a node v is calculated as the probability of a successful seeding of v, multiplied by the value of such an event:

$$U(v, t, \vec{X}(t)) = P(\vec{X}^v(t+1)) \cdot [1 + \sum_{w \in N_v^0(t+1)} U(w, t+1, \vec{X}^v(t+1))]$$

Note that the formulation of $U(v, t, \vec{X}(t))$ is recursive, and may involve successive iterations to evaluate the value of future seeding events beyond t+1. For practicality reasons, we limit the recursion to a depth of $k \in \{0, 1, 2\}$ iterations, as we found empirically that increasing the complexity of the algorithm by using higher k values has a diminishing return effect. The recursive computation of the score, for a depth of k iterations (k is provided as an input parameter), is shown in detailed in Algorithm 1.

Algorithm 1 The SSH Scoring Algorithm

Input:

t - time-step
$\vec{X}(t)$ - states of nodes in time-step t
v - node
k - recursion depth

Output:

Score of v

$P_v(t) \leftarrow P_v^{ind} + P_v^{soc} \cdot min(1, \frac{|N_v^1(t)|}{\theta_v})$
if $k = 0$ **then**
  $Score \leftarrow P_v(t)$
else
  $Score \leftarrow 1$
  for u in $N_v^0(t)$ **do**
    $Score \leftarrow Score +$ **SSH**$(t+1, \vec{X}^v(t+1), u, k-1)$
  end for
  $Score \leftarrow P_v(t) \cdot Score$
end if
return $Score$

To illustrate how Algorithm 1 works, consider the five-nodes network depicted in Figure 3.

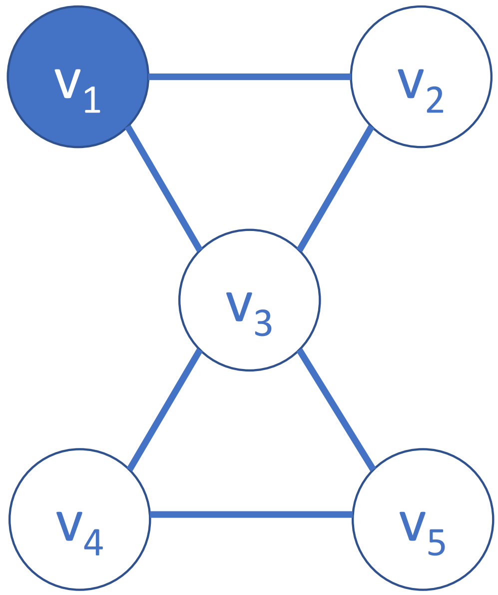


Figure 3: An illustration of a network with five nodes.

Assume that node v_1 was activated at the previous time-step $(t-1)$ and the following parameters: $P_v^{ind} = 0$, $P_v^{soc} = 1$ and $\theta_v = 0$, for all network nodes.

For a recursion depth of $k = 0$, we get the following utility scores for nodes v_2, v_3, v_4 and v_5 respectively:

$$SSH(t, \vec{X}(t), v_2, 0) = P_{v_2}(t) = 0 + 1 \cdot \frac{1}{2} = 0.5$$

$$SSH(t, \vec{X}(t), v_3, 0) = P_{v_3}(t) = 0 + 1 \cdot \frac{1}{2} = 0.5$$

$$SSH(t, \vec{X}(t), v_4, 0) = P_{v_4}(t) = 0 + 1 \cdot \frac{0}{2} = 0$$

$$SSH(t, \vec{X}(t), v_5, 0) = P_{v_5}(t) = 0 + 1 \cdot \frac{0}{2} = 0$$

Since both nodes v_2 and v_3 obtained the highest utility score, we will choose to seed either one of them at time-step t.

Alternatively, for a recursion depth of $k = 1$ we get the following utility scores for nodes v_2, v_3, v_4 and v_5 respectively:

$$SSH(t, \vec{X}(t), v_2, 1) = P_{v_2}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_3, 0)) = 0.5 \cdot (1 + 1) = 1$$

$$SSH(t, \vec{X}(t), v_3, 1) = P_{v_3}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_2, 0) + SSH(t+1, \vec{X}(t), v_4, 0) +$$
$$+ SSH(t+1, \vec{X}(t), v_5, 0)) = 0.5 \cdot (1 + 1 + 0.5 + 0.5) = 1.5$$
$$SSH(t, \vec{X}(t), v_4, 1) = P_{v_4}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_3, 0) + SSH(t+1, \vec{X}(t), v_5, 0)) = 0$$
$$SSH(t, \vec{X}(t), v_5, 1) = P_{v_4}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_3, 0) + SSH(t+1, \vec{X}(t), v_4, 0)) = 0$$

Since node v_3 obtained the highest utility score, we will choose to seed it at time-step t.

Runtime Complexity Analysis: The higher time-consuming operations of Algorithm 1 are performed in steps 1 and 6-8. In step 1, the algorithm determines the number of infected neighbors of node v, and in steps 6-8, the algorithm determines the utility score of each one of the non-infected neighbors of node v, given that v was already seeded successfully. Line 7, in particular, includes a recursive call which reduces the recursion depth (k) by 1. Therefore, if we denote the maximum degree of a node by d, the runtime complexity of Algorithm 1 in the worst case is $O(d^{k+1})$. It is important to note that Algorithm 1 is executed for each one of the non-infected nodes in the network, every time a seeding decision has to be made. Therefore, denoting the number of nodes in the network as $|V|$ and the seeding budget as B, the total time spent on Algorithm 1 is $O(|V| \cdot B \cdot d^{k+1})$.

5. Evaluation

In this section, we present an extensive set of empirical experiments that compare the performance of the proposed SSH approach (that is state-based) with that of existing seeding heuristics that rely on the network topology without taking into consideration the states of the nodes.

5.1. Experimental Setting

All the experiments were implemented in Python 2.7 and executed on a Linux machine running Centos 7.1, with 128 GB of RAM and a single Intel 2.7 GHz CPU.

Each of the simulations was preceded with selecting a random set of nodes, served as an initially infected population of size F. The infection time-steps of the nodes in this initial population were drawn uniformly from the interval $[-t_{inf}, -1]$. Then, at each time-step of the simulation, a single node was seeded, where the selection of the seeded node was based on different heuristics (the set of examined seeding heuristics is described below). Each seeding attempt either succeeded or failed in accordance with Eq. 1. The transitions in states of nodes were re-calculated at each discrete time-step.

The simulation ended when the entire budget of seeding attempts, B, was used. At this point, the final seeding success rate was calculated for each of the heuristics.

5.1.1. Parameters' Space

In the experiments, we examined a variety of values for the different parameters. In each set of simulations that are reported below, all parameters except one were set to their default value (fixed in most cases to the median of their examined range of values), while a single remaining parameter was examined over a varying range of values. The parameters' space used in our experiments is detailed in Table 1. Each combination of parameters values was examined by executing 400 simulation runs, for each one of the compared heuristics.

Table 1: Simulation Parameter Space

Parameter	**Values**
Network Topology (see Table 2)	**Sampled Citation network**, Slashdot network, Sampled EuEmail network, WikiVote network, Epinions network, Enron network
Network size (# of sampled nodes)	5000, 10000, 50000, **100000**, 500000, 1000000
Initially infected population size (F)	50, 100, **200**, 500, 1000
Budget (B)	50, 100, **200**, 500, 1000
Threshold (θ_v)	3, 4, **5**, 6, 7
Maximal Social Effect (P_v^{soc})	0.1, 0.3, **0.5**, 0.7, 0.9
Infection Time (t_{inf})	10, 20, **50**, 100, 200
Individual Effect (P_v^{ind})	**0**, 0.1, 0.2, 0.3, 0.4, 0.5

* The default value of each parameter is marked in **bold**.

In most of our experiments, we assumed that the values of the parameters θ_v and P_v^{soc} are known. In another dedicated experiment, we assumed that the distributions of these parameters' values are normal, and we only know their mean and standard deviation. These means are denoted by μ_θ and $\mu_{P^{Soc}}$, while the standard deviations are denoted by σ_θ and $\sigma_{P_v^{Soc}}$, respectively. The actual values of these parameters for each node, were randomly generated prior to each simulation run, and were not used in any way by the SSH heuristics.

5.1.2. Network Topologies

The simulations were executed on different network topologies, as detailed in Table 2. These topologies represent snapshots of real-world social networks, with some adaptations to our experimental framework, such as converting the networks to undirected, or sampling a subset of nodes. The original social network datasets are publicly available at (Leskovec & Krevl, 2014)

Table 2: Networks Used in Simulation

Network	**Number of Nodes**	**Average Degree**	**Average Clustering**	**Sampled?**
Citations	1000000	2.83	0.04	Yes
Citations	500000	4.06	0.06	Yes
Citations	100000	7.60	0.14	Yes
Citations	50000	8.20	0.16	Yes
Citations	10000	6.81	0.20	Yes
Enron	36692	10.02	0.50	No
WikiVote	7115	28.32	0.14	No
Slashdot	82168	14.18	0.06	No
EuEmail	100000	1.57	0.03	Yes
Epinions	75879	10.70	0.14	No

5.1.3. Seeding Heuristics

We compared three variations of the proposed SSH approach (SSH-0, SSH-1 and SSH-2, where the levels of recursion were $k = 0$, $k = 1$ and $k = 2$ respectively) with four benchmark approaches as we proceed to describe. These benchmark approaches included both a state-of-the-art network-centrality-based approach (GEC), and a simple random selection of nodes (Random).

Furthermore, for each of these two benchmark approaches we added a variation which considered as optional seeding candidates, only nodes that have a non-zero probability to become infected (i.e., nodes that have at least one infected neighbor). These additional variations were named Picky-GEC and Picky-Random.

The seven heuristics mentioned above are described in further details below:

Random Randomly seeds one uninfected node at each time-step.

GEC Chooses the uninfected node with the highest Eigenvector Centrality measure at each time-step.

Picky-Random Randomly chooses an uninfected node from the nodes that have a non-zero probability to become infected.

Picky-GEC Chooses the uninfected node with the highest Eigenvector Centrality from the nodes that have a non-zero probability to become infected.

SSH-0 - Chooses the uninfected node with the highest value of $P_v(t)$ at each time-step (i.e., Algorithm 1 with k = 0).

SSH-1 - Chooses the uninfected node with the highest value of $P_v(t)$ at each time-step (i.e., Algorithm 1 with k = 1).

SSH-2 - Chooses the uninfected node with the highest value of $P_v(t)$ at each time-step (i.e., Algorithm 1 with k = 2).

5.2. Results

5.2.1. Overall Comparison of SSH with the other Benchmark Methods

Figure 4 presents an overall comparison of the SSH approach to the other benchmark methods. Figure 4 (top) presents this comparison for different network topologies whereas Figure 4 (bottom) focuses on different sample sizes of the Citation network topology. In these experiment, all other parameters that are mentioned in Table 1 except for the network topology and size were set to their default values.

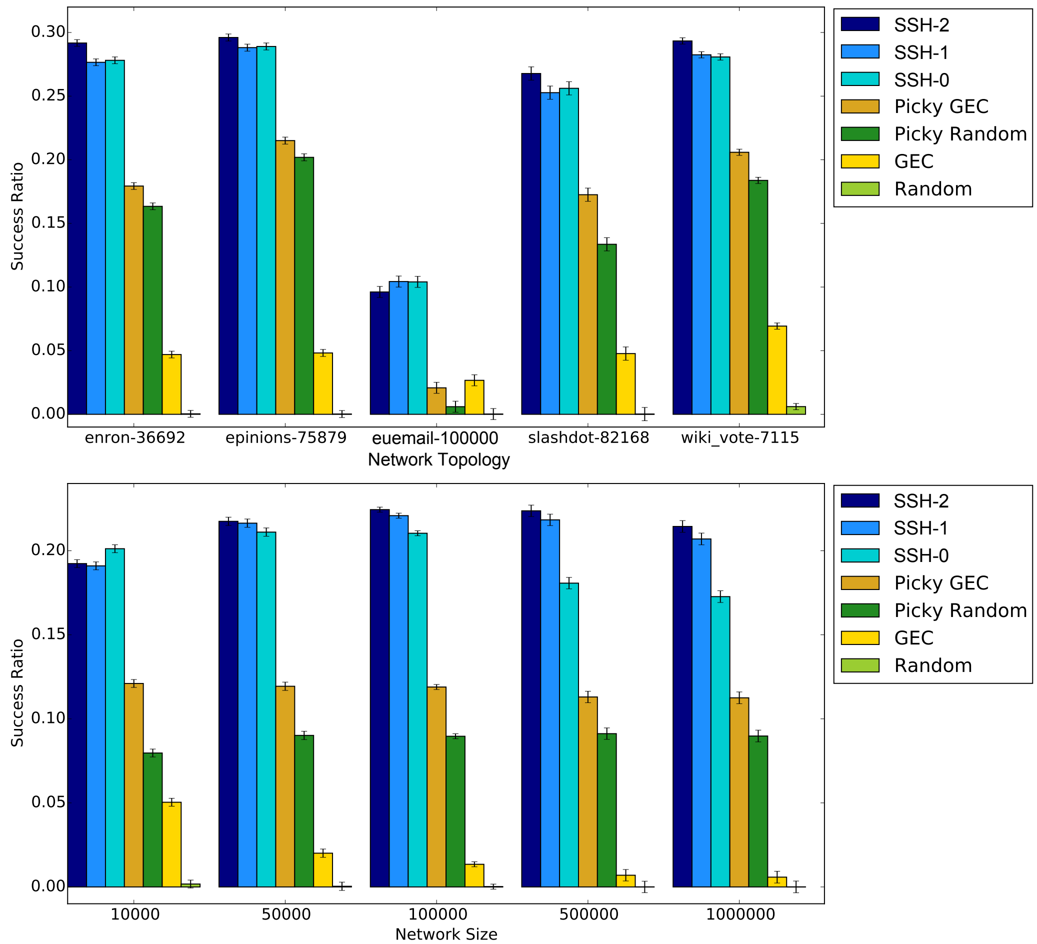


Figure 4: An overall comparison of SSH (the three blue bars) with the benchmark methods, across different network topologies (top) and network sizes (bottom).

As can be seen in the figure, the three SSH heuristics (blue bars) significantly outperform the other benchmark methods. More specifically, comparing SSH-0 (the worst out of the three SSH heuristics) to Picky-GEC (the best out of the other benchmark methods), the improvement ranges from 30% to 75%.

With regard to the different SSH heuristics, it seems that in most cases SSH-2 achieves the best performance, followed by SSH-1 and then SSH-0. This is in accordance with the amount of information that each of those heuristics uses to evaluate the scores of potential nodes to seed. However, it is worth mentioning that the differences in performance between these three heuristics are relatively low in comparison to the other benchmark methods.

As expected, the worst performing heuristic (by far) is the Random heuristic, which does not utilize any information about the network topology nor the states of the nodes. The GEC heuristic, performs slightly better than Random heuristic, since it utilizes information about the network topology.

Two interesting heuristics are Picky-Random and Picky-GEC that utilize partial information about the states of the nodes (i.e., which nodes have non-zero probability to be seeded successfully). As can be seen in the figure, these two heuristics perform better than the basic Random and GEC heuristics but worse than the SSH heuristics. We can also see that Picky-GEC performs slightly better than Picky-Random since it also utilizes information on the network topology.

5.2.2. Centrality of Seeded Nodes

In the previous experiment, we saw that the SSH heuristics perform significantly better than the GEC heuristic. In order to understand better why this is the case, we compared the centrality of nodes that were chosen by each of the two approaches. We were mainly interested to know if the SSH heuristics select to seed central nodes, or if it chooses to seed less central nodes. Note that in real-world marketing scenarios that involve seeding, not all seeding actions have the same cost. In fact, highly central nodes in social networks often represent celebrities, and the cost of seeding such celebrities is likely to be higher than that of less known individuals. Figure 5 presents the Eigenvector centrality of nodes that were chosen for seeding by the SSH-1 and GEC heuristics, along time.

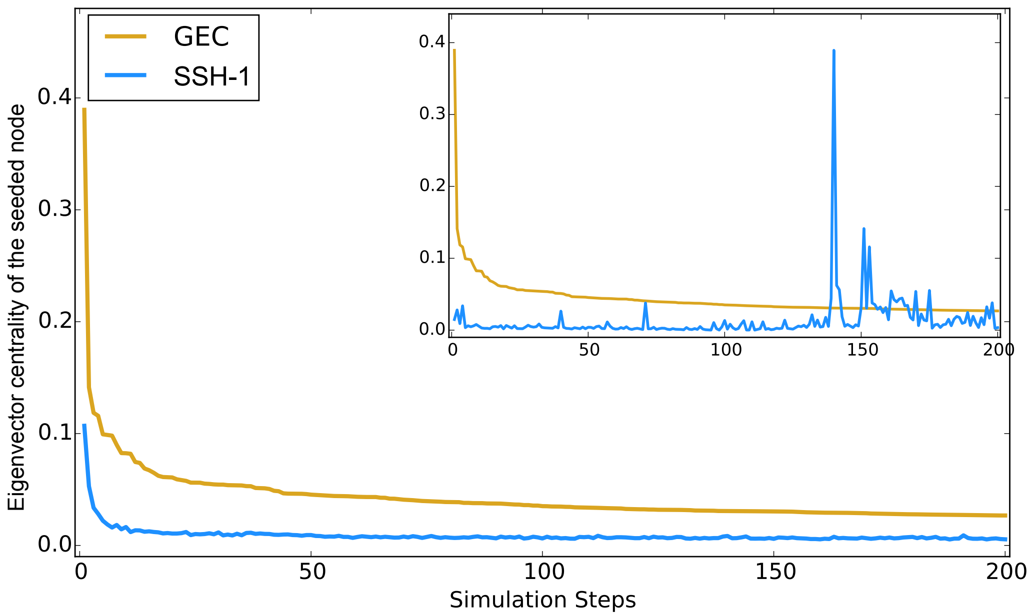


Figure 5: Eigenvector centrality of the nodes chosen for seeding along time.

The exterior figure presents the average Eigenvector centrality of nodes over 400 executions, where all parameters from Table 1 were set to their default values. As can be seen from the figure, both heuristics tend to start with nodes that have a higher Eigenvector centrality score and continue with nodes with lower and lower Eigenvector centrality scores. While this observation is expected for the GEC heuristic, it is less expected for the SSH-1 heuristic, since it does not make an explicit use of the network topology. It can also be seen that the average centrality score of the nodes selected by the SSH-1 heuristic is substantially lower than that of the GEC heuristic.

The interior figure presents a single execution, out of these 400 executions, for each of the two heuristics. As expected, the GEC heuristic performs the same in the single execution case and in the average case. However, with regard to the SSH-1 heuristic, we notice that central nodes are chosen somewhere at the middle of the contagion process and not necessarily at the initial stages. In other words, at any given time, the SSH-1 heuristic might prefer to choose a non-central node over a central node as long as its expected utility (its likelihood to be seeded successfully and its impact on its neighbors) is considered higher. This observation, together with the superiority of the SSH approach (as demonstrated in the previous experiment), emphasize the importance of utilizing the states of the nodes and not only the network topology when assessing their ability to spread information. This is especially interesting since, centrality measures of a node, such as Eigenvector centrality, which take into account the network topology only, are often considered in the literature as a good proxy for the node's ability to spread information.

5.2.3. Sensitivity Analysis of the Model's Parameters

Figure 6 shows the total number of successful seeding attempts as a function of the seeding budget B. As expected, the number of successful seeding attempts grows with the budget size for all heuristics, but this growth presents a "diminishing return" effect. The figure also demonstrates the superiority of the SSH approach (blue plots), where its gap from the other heuristics increases with the budget size.

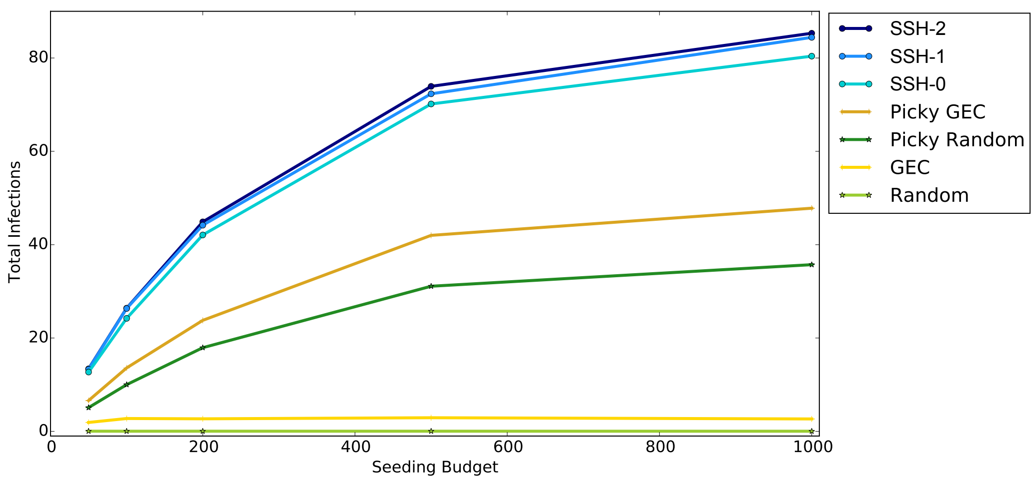


Figure 6: The number of successful seeding attempts as a function of the seeding budget B.

As described in the previous section, we assume the existence of an initially infected population of size F, prior to the beginning of the seeding attempts. Figure 7 reports the influence of F on the success rate of the different seeding heuristics. As expected, larger F values lead to higher success rates for all of the heuristics. While this increase exists, but is barely noticeable for the Random and Picky Random heuristics, it is clearly evident in the case of the SSH heuristics. Here as well, the SSH heuristics outperform the other heuristics, even for small values of F, and the gap becomes larger as F grows.

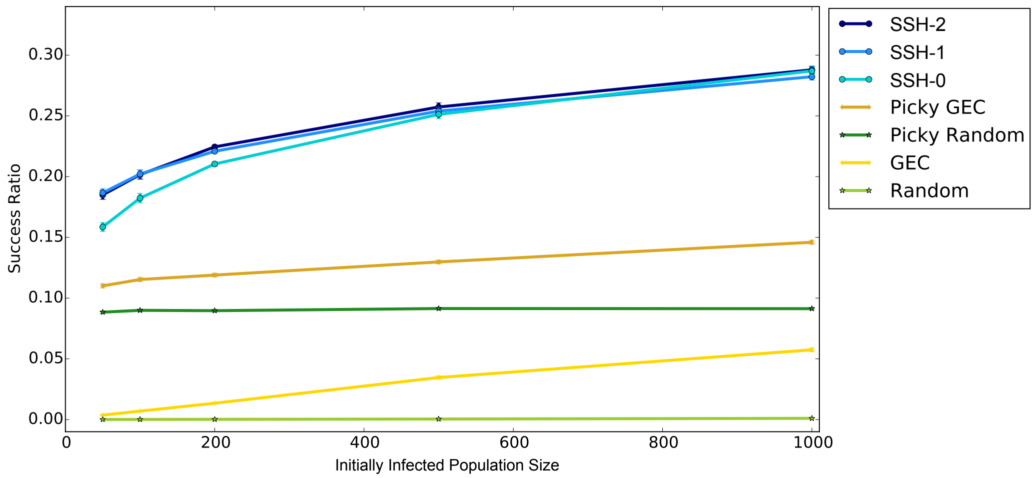


Figure 7: The proportion of successful seeding attempts as a function of the initially infected populations size F.

Figure 8 reports the influence of the infection time t_{inf} on the success rate of the different seeding heuristics. As can be seen in the figure, larger t_{inf} values lead to higher success rates for all of the heuristics. This is quite expected since lower t_{inf} values imply shorter infectious period of newly infected nodes, leading to lower social influence in the network at any given time. When the infection time is significantly short (around 5-10 time-steps), all of the heuristics suffer from poor performance. However, infection times of 50 time-steps and above result in high performance, where the improvement in performance gradually decreases with higher values of t_{inf}. Again, we see that the SSH approach (blue plots) significantly outperforms the other heuristics, for all of the examined values of t_{inf}.

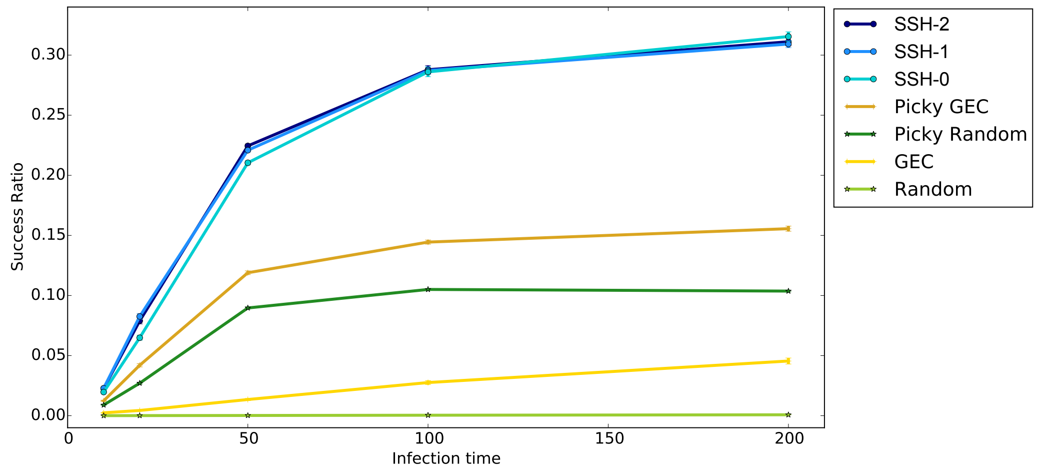


Figure 8: The proportion of successful seeding attempts as a function of the infection time t_{inf}.

The effect of the maximal social effect p_v^{soc} and the social threshold θ_v on the success rate of the different seeding heuristics is demonstrated in Figure 9. As can be seen in Figure 9 (top), higher values of p_v^{soc} are associated with higher success rates for all heuristics, as expected. Interestingly, the SSH approach grow super-linearly with p_v^{soc}, whereas all other heuristics grow roughly linearly. This causes the gap between the SSH approach (blue plots) and the other heuristics to become larger with higher values of p_v^{soc}. Indeed, when the social forces are stronger, the SSH approach, which better utilizes the information about the social influence is expected to reach better results. A similar (though inversed) trend of what was observed in Figure 9 (top) is presented in Figure 9 (bottom). This inversed trend is quite expected due to the $\frac{P_v^{soc}}{\theta_v}$ element in Eq. 1.

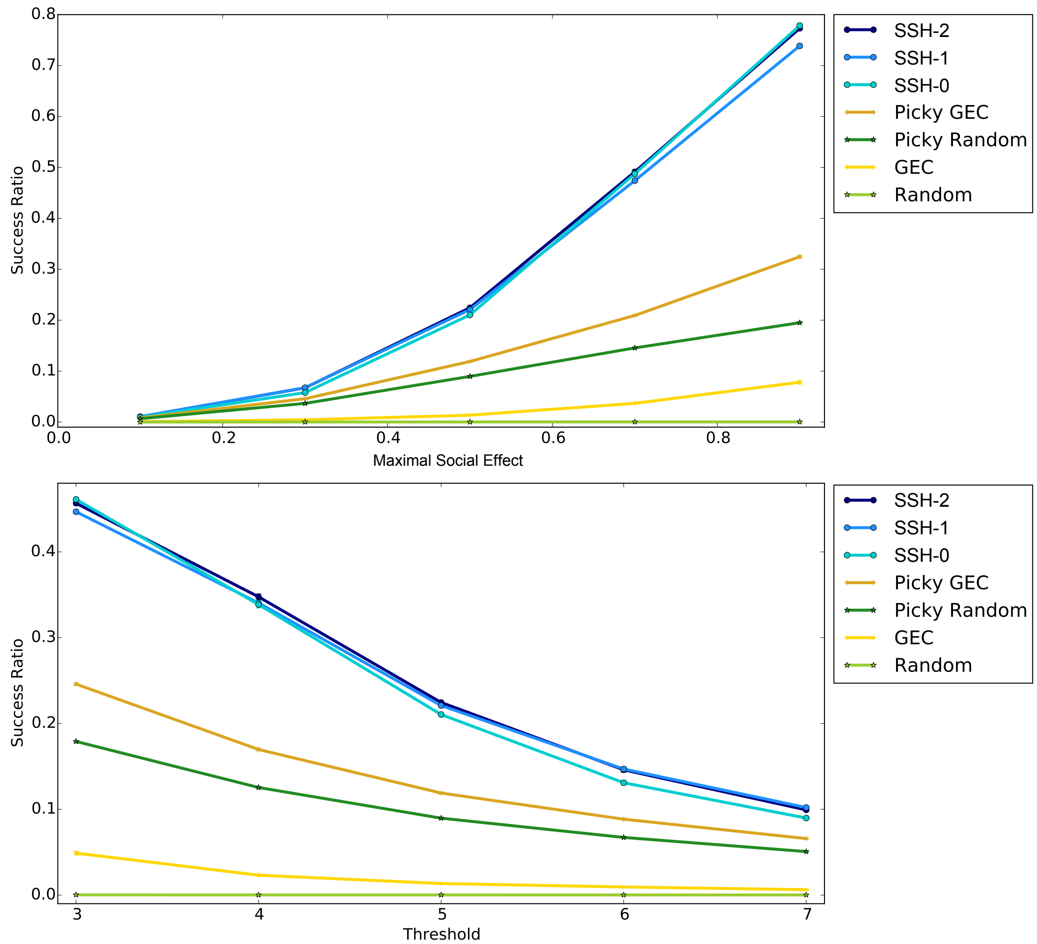


Figure 9: The proportion of successful seeding attempts as a function of the maximal social effect P_v^{soc} (top) and the social threshold θ_v (bottom).

While all of the above analyses focused on the social effect, where we set the individual effect to $P_v^{ind} = 0$, we now turn to analyzing the effect of the individual (non-social) effect on the success rate of the different seeding heuristics (see Figure 10). First, we observe that the success rates of all seeding heuristics increase with the individual effect P_v^{ind}. We can also see that the growth rate is similar in all heuristics, including the Random heuristic. This observation makes sense, since large values of P_v^{ind}, significantly reduce the importance of the social effect, and therefore make the scheduled approach less necessary. Similarly, we also see that for larger values of individual effect (i.e., $P_v^{ind} \geq 0.05$), the SSH-0 heuristic outperforms the SSH-1 and SSH-2 heuristics.

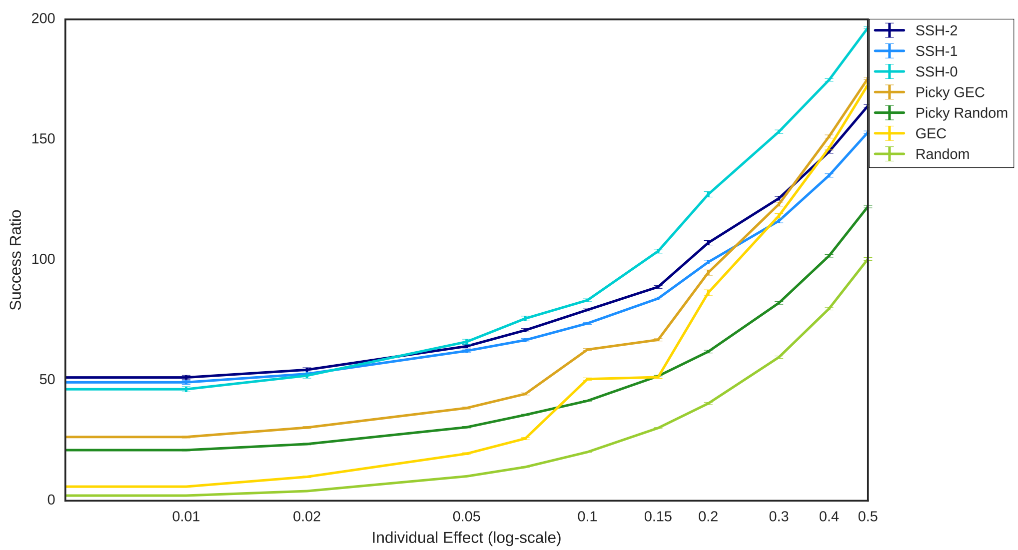


Figure 10: The number of successful seeding attempts as a function of the individual effect P_v^{ind}.

5.2.4. Introducing Uncertainty

The results described in the section above were obtained by assuming that the values of P_v^{soc} and θ_v are known. In most cases however, this is not a realistic assumption. At best, the distribution of these parameters can be estimated from previous marketing campaigns, but the specific parameter value for each person is still considered unknown. Based on this understanding, we conducted another experiment to inspect the performance of the proposed SSH approach within a more realistic scenario, in which P_v^{soc} and θ_v are assumed to be normally distributed and their means and standard deviations are assumed to be known; however, the actual values for each node are considered unknown.

Accordingly, in each set of executions, we first chose the mean and standard deviation. Then, we generated the "real" values for P_v^{soc} and for θ_v for each node based on the chosen distributions. Finally, we ran the different seeding heuristics where the means of the distributions were given as inputs, instead of their actual values. Note that in these experiments, the real values are only used in the simulative process, but is not used by the seeding node selection process.

Figure 11 reports the success rate of the different heuristics as a function of uncertainty (reflected by SD/mean).

The interior figure shows the success rate of Picky-Random as a function of uncertainty. As can be seen from the figure, the success rate increases moderately with uncertainty. The explanation for this is that high uncertainty values lead to a larger number of nodes with high P_v values (due to high P_v^{soc} and low θ_v values).

The exterior figure reports the relative success rate of the different heuristics, normalized with respect to Picky-Random, as a function of uncertainty. As can be seen from the figure, while the GEC, Picky-GEC and Random heuristics preserve the same relative success rate when uncertainty increases, the success rate of the SSH approach decreases. This is quite expected, as the SSH approach explicitly relies on the values of P_v^{soc} and θ_v for calculating the scores of nodes. Thus, an inaccurate estimation of these values due to a large standard deviation, leads to poorer selection of nodes and to a reduced performance. In contrast, all other heuristics which do not rely on the values of P_v^{soc} and θ_v, and therefore are not affected by inaccurate values of P_v^{soc} and θ_v. Nevertheless, even in relatively high uncertainty levels, the success rate of the SSH approach is still significantly higher than that of the other methods.

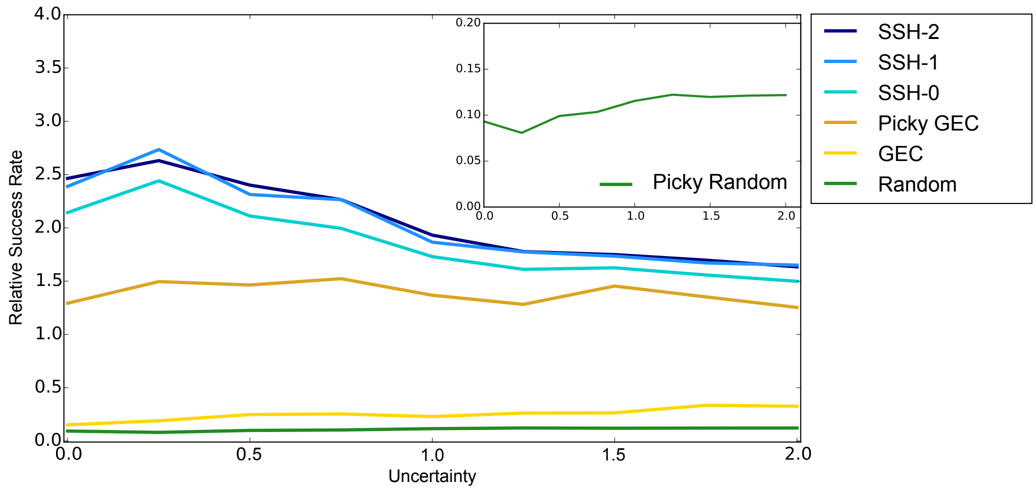


Figure 11: The improvement rate as a function of the degree of uncertainty (measured as the standard deviation of P_v^{soc} and θ_v).

5.2.5. Runtime

The different SSH heuristics represent growing degrees of future planning effort. While SSH-0 is fully greedy, in terms of planning only the current step, SSH-1 tries to plan one step ahead, and SSH-2 method plans two steps ahead. Although the SSH approach can be used with even higher number of planning steps (i.e., higher than 2), we did not find such large number of planning steps more effective. This observation is of high importance since the computational cost of planning ahead significantly increases with the network size, and due to the tremendous sizes of real-world social networks.

Figure 12 reports the runtime of the different heuristics as a function of the network size (different sample sizes of the Citation network). The runtime of SSH-0 and Picky-Random are roughly the same since they require to perform $O(1)$ operations for each one of the network nodes in each iteration. The runtime of SSH-1 is slightly higher since it requires some calculations of the first social circle of each network node in each iteration. The runtime of SSH-2 is again significantly higher than the runtime of SSH-1, since it requires some calculations on the first and second social circles of each network node (which cover a large fraction of the entire network) in each iteration. The runtime of GEC and Picky-GEC is also very high since it requires to calculate the Eigenvector centrality score for each of the network nodes (this is done once for each node, but the calculation is still expensive). Finally, we observe that starting from a certain network size (700,000), the runtime of GEC and Picky-GEC becomes even higher than that of SSH-2. Since the run-times of SSH-0 and SSH-1 seems to be reasonable, and since their success rate is almost as good as that of SSH-2, we will probably prefer to use them in future applications of real-world scenarios that involve large-scale networks.

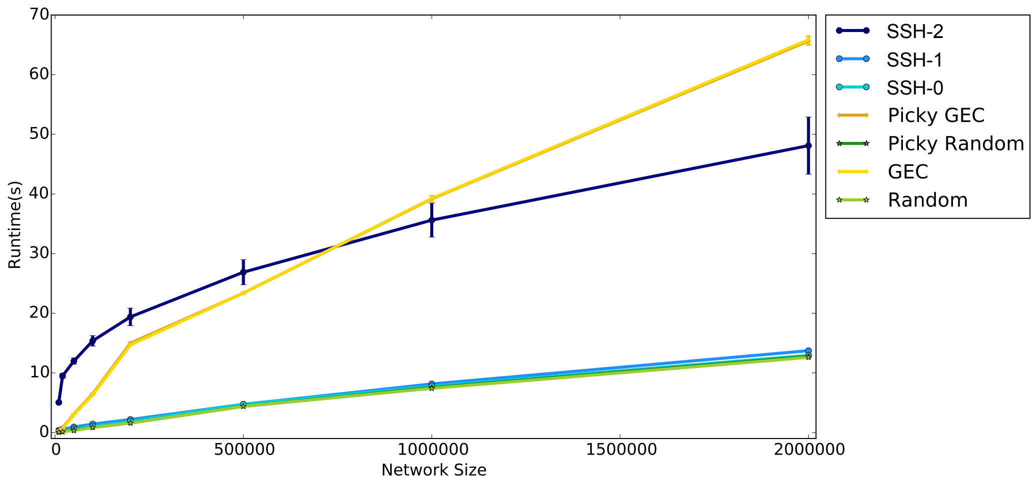


Figure 12: Runtime of the different heuristics as a function of the network size.

6. Summary and Future Work

Many works that study information diffusion in social networks consider a phenomenon by which information spreads virally through the network. Yet, unlike the spread of biological viruses that can be carried passively by agents and infect a significant portion of the network, information cascades are known to be shorter while long cascades are rather rare. These results do not necessarily imply that social impacts lost their importance, but rather that people spread information in a more selective way, which does not necessarily fit the assumptions of traditional models of infectious diseases.

We propose a new information diffusion model, named Active Viral Marketing (AVM), in which agents, e.g., sales representative of a company, communicate with network users, e.g., potential clients, and offer them a new product or service. The probability that a user accepts such an offer is based on the previous adoption rate of his/her friends, as well as his/her own tendency toward the product.

Since promotion actions often incur some financial cost (limiting the number of clients that can be approached), the company has to select which users to approach and at what time, in order to increase the total adoption rate in the network. The need for a correct timing of approaching a customer is a direct result of memory retention loss, where new products quickly become an old habit and therefore the likelihood of influencing a peer node to purchase the new product quickly decays. The proposed Scheduled Seeding Heuristics (SSH) for user selection, chooses nodes that are most likely to accept an offer at any given time-step, and thus are more likely to influence their own non-infected neighbors at the next time-step.

In a large set of simulations, we show that the proposed heuristics increase the adoption rate in 30%-75% (depending on the initial conditions), over a state-of-the-art method that seeds the nodes according to their Eigenvector centrality score.

Having indicated that, it is important to note that the proposed method is mainly applicable to products that have a viral characteristic. These are products or services where a substantial part of the purchasing decision is based on social influence. In products or services for which the social forces are significantly less influential, it might be better to use the existing state-of-the-art methods of selecting nodes based on the network's topological properties.

Most diffusion models, including the proposed model, assume that all seeding actions have the same cost. As mentioned in Section 5.2.2, highly central nodes in social networks often represent celebrities or influencers, and the cost of seeding such entities is likely to be higher than that of less known individuals. Future studies should take into account different seeding costs for different nodes, depending for example on the network topology.

An interesting future extension to this work would be to study diffusion models that combine both the traditional passive infection together with the proposed continuous active seeding. Such a combined model is expected to be applicable for a wider range of real-world scenarios than each one of the two isolated models. Furthermore, it would be interesting to extend the proposed utility-based heuristics to support such a combined model.

The evaluation of this study is mainly based on simulations that utilize real-world network topologies. In future works, it would be interesting to enrich these simulations with additional real-world data such as purchasing history of users. In addition, it would be insightful to conduct a live experiment to compare the adoption rate obtained by the scheduled seeding approach versus the non-scheduled seeding approach.

Acknowledgements

This work was funded by the Kamin grant of the Israeli Chief Scientist (file number 58073).

Appendix A. Properties of the Influence Maximization Problem under the Active Viral Marketing Diffusion model

Appendix A.1. NP-Hardness

Claim: The influence maximization problem is NP-hard for the Active Viral Marketing diffusion model.

Proof: Consider an instance of the NP-hard Set Cover problem Garey & Johnson (1979): Given a collection of subsets $\{S_1, S_2, ..., S_m\}$ of a ground set $U = \{u_1, u_2, ..., u_n\}$, we wish to know whether there exist k of the subsets whose union is equal to U. We show that this can be viewed as a special case of the influence maximization problem for the Active Viral Marketing diffusion model. (We can assume that $k < n < m$.)

Given an arbitrary instance of the Set Cover problem, we define a corresponding directed graph as follows. The graph contains $1 + m + n$ nodes: a single node A, a node v_{S_i} for each subset S_i, a node v_{u_j} for each element u_j, and $m + \sum_{S_i} |S_i|$ directed edges: a directed edge (A, v_{S_i}) from A to each one of the v_{S_i} nodes and a directed edge (v_{S_i}, v_{u_j}) whenever $u_j \in S_i$.

In addition, consider the following parameters: $\theta = 1$, $t_{inf} = k$, $P_{ind} = 1$ and $P_{soc} = 0$ for node A, $\theta = 1$, $t_{inf} = 1 + k + n$, $P_{ind} = 0$ and $P_{soc} = 1$ for all other nodes, and a seeding budget of size $B = 1 + k + n$.

We note the following:

1. For the instance we have defined, activation is a deterministic process, as all probabilities are either 0 or 1.
2. A solution to the influence maximization problem must choose to seed node A at time-step $t = 0$ (seeding the node A at time-step $t = 0$ is assured to succeed while trying to seed any other node is assured to fail).
3. At least k out of the m nodes of type v_{S_i} must be seeded (a direct result of the seeding budget size).
4. Assuming that node A was seeded at time-step $t = 0$, seeding each one of the v_{S_i} nodes at time-steps $1 \leq t \leq k$ is assured to succeed (they only need one infected neighbor for the seeding action to succeed). Similarly, seeding each one of the v_{S_i} nodes at time-steps $t > k$ is assured to fail ($t_{inf} = k$ for node A).

5. Following the four bullet points above, it stems that a solution to the influence maximization problem must choose to seed node A at time-step $t = 0$, k out of the m nodes of type v_{S_i} at time-steps $1 \leq t \leq k$ and all of the n nodes of type v_{u_j} at time-steps $k+1 \leq t \leq k+n$.
6. Assuming that node A was seeded at time-step $t = 0$ and k out of the m nodes of type v_{S_i} were seeded at time-steps $1 \leq t \leq k$, seeding a node v_{u_j} at time-steps $k+1 \leq t \leq k+n$ will succeed only if there exists a node v_{S_i} for which $u_j \in S_i$ and v_{S_i} is one of the k chosen nodes at time-steps $1 \leq t \leq k$.
7. The maximum number of nodes that can be seeded successfully is $1 + k + n$ (due to the budget size).

The answer to the Set Cover problem is True if and only if the solution to the corresponding influence maximization problem led to the successful seeding of exactly $1 + k + n$ nodes. ($1 + k + n$ successful seedings mean that we managed to seed successfully node A, k out of the m nodes of type v_{S_i} and all n nodes of type u_j, which further imply that there exists k subsets that cover the entire set U).

Since the Set Cover problem is known to be NP-hard, then so is the influence maximization problem for the Active Viral Marketing diffusion model.

Appendix A.2. Sub-Modularity

Consider the Active Viral Marketing diffusion model defined above and the function F, which receives an ordered subset of network nodes to be seeded (at consecutive time-steps) as input, and returns the expected number of successful seedings as output. By definition, F is not sub-modular, since sub-modular functions receive a set rather than an ordered set as input. Moreover, even if we extend the definition of sub-modular functions to the case of ordered sets, F would still not satisfy the sub-modularity condition. To illustrate why, consider a network composed of two nodes v_1 and v_2 and a single edge between them, and the following parameters: $P_v^{ind} = 0.1$, $P_v^{soc} = 0.9$, $\theta_v = 1$ and $t_{inf} = 2$, for all network nodes. Now, consider the two ordered sets $X = ()$ and $Y = (v_1)$. The sub-modularity condition requires (among the rest) that adding v_2 to Y will result in a lower gain in F than adding it to X (since $X \subset Y$). More specifically, it is required that $F((v_1, v_2)) - F((v_1)) < F((v_2)) - F(())$ However, it is easy to see that $F(()) = 0$, $F((v_1)) = 0.1$, $F((v_2)) = 0.1$, and $F((v_1, v_2)) = 0.1 + (0.1 \cdot 1 + 0.9 \cdot 0.1) = 0.29$. Therefore, $F((v_1, v_2)) - F((v_1)) = 0.19 > F((v_2)) - F(()) = 0.1$ and the sub-modularity condition is violated.

References

Anderson, R. M., May, R. M., & Anderson, B. (1992). *Infectious diseases of humans: dynamics and control* volume 28. Wiley Online Library.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men. S*, (pp. 222–236).

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *Science*, *341*, 1236498.

Barthélemy, M., Barrat, A., Pastor-Satorras, R., & Vespignani, A. (2004). Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical Review Letters*, *92*, 178701.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social networks*, *29*, 555–564.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, *27*, 55–71.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, *25*, 163–177.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American journal of Sociology*, *113*, 702–734.

Chen, W., Wang, C., & Wang, Y. (2010a). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1029–1038). ACM.

Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 199–208). ACM.

Chen, W., Yuan, Y., & Zhang, L. (2010b). Scalable influence maximization in social networks under the linear threshold model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 88–97). IEEE.

Chierichetti, F., Kleinberg, J., & Panconesi, A. (2014). How to schedule a cascade in an arbitrary graph. *SIAM Journal on Computing*, *43*, 1906–1920.

Fibich, G. (2016). Bass-sir model for diffusion of new products in social networks. *Physical Review E*, *94*, 032305.

Garey, M. R., & Johnson, D. S. (1979). Computers and intractability: A guide to the theory of npcompleteness (series of books in the mathematical sciences), ed. *Computers and Intractability*, (p. 340).

Goel, S., Watts, D. J., & Goldstein, D. G. (2012). The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce* (pp. 623–638). ACM.

Goldenberg, J., Libai, B., & Muller, E. (2001a). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, *12*, 211–223.

Goldenberg, J., Libai, B., & Muller, E. (2001b). Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, *2001*, 1.

Goldfarb, A., & Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, *30*, 389–404.

Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 241–250). ACM.

Goyal, A., Bonchi, F., Lakshmanan, L. V., & Venkatasubramanian, S. (2013). On minimizing budget and time in influence propagation over social networks. *Social network analysis and mining*, *3*, 179–192.

Goyal, A., Lu, W., & Lakshmanan, L. V. (2011). Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web* (pp. 47–48). ACM.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*, 1420–1443.

Hinz, O., Skiera, B., Barrot, C., & Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, *75*, 55–71.

Jankowski, J., Bródka, P., Kazienko, P., Szymanski, B. K., Michalski, R., & Kajdanowicz, T. (2017a). Balancing speed and coverage by sequential seeding in complex networks. *Scientific reports*, *7*, 891.

Jankowski, J., Bródka, P., Michalski, R., & Kazienko, P. (2017b). Seeds buffering for information spreading processes. In *International Conference on Social Informatics* (pp. 628–641). Springer.

Jung, K., Heo, W., & Chen, W. (2012). Irie: Scalable and robust influence maximization in social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on* (pp. 918–923). IEEE.

Katz, E., & Lazarsfeld, P. (1955). Personal influence.

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 137–146). ACM.

Khelil, A., Becker, C., Tian, J., & Rothermel, K. (2002). An epidemic model for information diffusion in manets. In *Proceedings of the 5th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems* (pp. 54–60). ACM.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 497–506). ACM.

Leskovec, J., & Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web* (pp. 915–924). ACM.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007a). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 420–429). ACM.

Leskovec, J., & Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., & Hurst, M. (2007b). Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining* (pp. 551–556). SIAM.

Lin, S., Hu, Q., Wang, F., & Philip, S. Y. (2014). Steering information diffusion dynamically against user attention limitation. In *Data Mining (ICDM), 2014 IEEE International Conference on* (pp. 330–339). IEEE.

Long, C., & Wong, R. C.-W. (2011). Minimizing seed set for viral marketing. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on* (pp. 427–436). IEEE.

Mahajan, V., Muller, E., & Bass, F. M. (1991). New product diffusion models in marketing: A review and directions for research. In *Diffusion of technologies and social behavior* (pp. 125–177). Springer.

Newman, M. (). Networks: an introduction. 2010. *United Slates: Oxford University Press Inc., New York*, (pp. 1–2).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.*. Technical Report Stanford InfoLab.

Seeman, L., & Singer, Y. (2013). Adaptive seeding in social networks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on* (pp. 459–468). IEEE.

Tong, G., Wu, W., Tang, S., & Du, D.-Z. (2017). Adaptive influence maximization in dynamic social networks. *IEEE/ACM Transactions on Networking (TON)*, *25*, 112–125.

Vespignani, A. (2012). Modelling dynamical processes in complex sociotechnical systems. *Nature physics*, *8*.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific reports*, *2*.

Zhou, J., Liu, Z., & Li, B. (2007). Influence of network structure on rumor propagation. *Physics Letters A*, *368*, 458–463.

Examining the relationship between social media analytics practices and business performance in the Indian retail industry: The mediating role of customer engagement

Poonam Garg
Institute of Management Technology (IMT)
Ghaziabad, India
Pgarg@imt.edu

Bhumika Gupta*
LITEM Laboratory, Institut Mines-Telecom Business School,
Evry, France
Bhumika.gupta@imt-bs.eu

Sam Dzever
LITEM Laboratory, Institut Mines-Telecom Business School,
Evry, France
Sam.dzever@imt-bs.eu

Uthayasankar Sivarajah
University of Bradford, School of Management
Bradford, United Kingdom
U.sivarajah@bradford.ac.uk

Vikas Kumar
Bristol Business School, University of the West of England,
Bristol, UK
Vikas.kumar@uwe.ac.uk

*Corresponding Author

Examining the Relationship between Social Media Analytics Practices and Business Performance in the Indian Retail and IT Industries: The Mediation Role of Customer Engagement

Abstract

Social media analytics (SMA) is a dynamic field which has received considerable attention from both academics and management practitioners alike. A significant number of the scholarly research currently being conducted in SMA, however, is conceptual. Industry experts know that SMA creates new opportunities for organisations who want to more strongly engage with their customers and improve business performance. However, the relationship between social media analytic practices (SMAP), customer engagement (CE), and business performance (BP) has not yet been sufficiently investigated from an empirical perspective. In order to gain a better understanding of the relationship between SMAP and BP and the mediation role of CE in that process, a large-scale survey was conducted among senior and mid-level managers as well as consultants in the Retail and information technology (IT) industries in India. Specifically, a structured closed-ended questionnaire was administered to managers and management consultants country-wide and gathered usable responses from 281 respondents holding positions such as: Digital Marketing Executive/Digital Marketing Specialist, Management Consultant, Analytics Manager, Customer Relationship Manager, Marketing Director, Engagement Manager, etc. who were in charge of digital marketing strategies in the respondent retail and IT organisations. The questionnaire addressed issues related to the way in which SMAP contribute to an enhanced business performance through the mediation role of customer engagement. Structural Equation Modelling was employed to analyse the received empirical data. On the basis of the findings our research concludes that there is a significant positive relationship between SMAP and BP mediated by CE in the Indian retail and IT industries.

Keywords: *Social Media Analytics, Customer Engagement, Business Performance, Indian Retail and IT Industries.*

1. Introduction

Social media, a powerful 21st century communication medium, has changed the dynamics of the business environment and redefined the way organisations communicate and engage with each other and their stakeholders. It has also, at the same time, provided the opportunity for customers to share their experiences about a product or brand. As a result, it is vital for businesses to be aware of the perception customers have of their products and/or brand identity (Anjanita, 2017). Data derived from Statista Database Company (2019) suggest that Facebook is one of the world's most popular social networking sites with nearly 2.2 billon active users. From the consumers' perspective, social media and social networks have become an essential part of their daily lives (Shiau *et al*, 2017; Shiau *et al*, 2018) and this, in turn, has changed the way in which individual consumers acquire information and communicate with each other (Dwivedi *et al*, 2015). Presently, Instagram is the most popular photo and video-sharing platform and enjoys one billion monthly active accounts (Clement, J., 2019). For its part, with 330 million monthly active users Twitter generates 6,000 tweets per second (Statista Database Company, 2019; Internet lives today, 2018). From social media data, retail and IT organisations have demonstrated a keen interest in analysing, measuring, and predicting business/customer insights in order to make sound business decisions (Sivarajah *et al*. 2017; Rafiq, 2017). According to Sivarajah *et al* (2019), SMA has the potential to help businesses understand the effectiveness of organisational communication and customer interaction on different social media platforms. SMA has been referred to as an approach to gathering data from social media networking sites and blogs then analysing it according to online activities of customers, user-generated data, customer sentiments and customer behaviour in real time to enable more efficient and effective business decisions to be made (Bekmamedova & Shanks, 2014).

SMA is currently helping retailers to collect and analyse information in a way that enables them better understand customer behaviour, enhance customer life cycle, engage new markets, improve responsiveness, and inspire loyalty. Kapoor *et al* (2018) carried out a review of social media and social networking studies from 1997 to 2017. Their study identifies the major advances made in social media research during the period and highlights their significance. A similar study by Stieglitz *et al* (2018) outlines the key challenges and steps in the SMA process and their mitigating strategies. A number of success stories were reported from consultancies and commercial research companies in the form of "white papers" which highlight the use of SMA in enhancing customer engagement and business performance (SAS, 2011; The Enterprise Strategy Group, 2013; Cognizant, 2014). Customer engagement relates primarily to the nature and intensity of a relationship a customer has with the supplier of a product or service; while business performance refers primarily to the extent to which such relationship contributes to the organisation's success and prosperity.

India is today a very fast growing economy with a rapidly expanding retail and IT industries. Retailers such as Reliance Retail, RPG Retail, Future Group, Aditya Birla Group, ITC Ltd, Tata Group, Vishal Group, and BPCL, and many multinational companies such as Wal-Mart, Tesco, and Metro have established in India and are rapidly strengthening their market positions. According to India Brand Equity Foundation (IBEF), the Indian retail market is expected to increase by 60% to reach US$ 1.1 trillion by 2020; and the online retail market is expected to grow from US$ 17.8 billion in 2017 to US$73 billion by 2022 (IBEF, 2018). According to Ernst & Young (EY) and Retailers Association of India (RAI), "Organised retail penetration, currently estimated at 7.5%, is expected to reach 19-20% p.a. growth… by 2018" (EY and RAI, 2013). Research shows that there is a lack of studies that have explored the application of social media analytics on the Indian retail sector particularly exploring the relationship between SMA practices, customer engagement and business performance. This study therefore attempts to fill this gap in research gap.

The objective of this research, therefore, is to empirically investigate and produce knowledge about the nature of the relationship that exists between customer engagement and business performance. We use SMAP as the tool for our investigation. Four research questions have been identified in our research as follows:

(1) Does strategic use of SMAP have a positive relationship with customer engagement?
(2) Does customer engagement have a positive relationship with business performance?
(3) Does strategic use of SMAP have a positive relationship with business performance?
(4) Does customer engagement have a mediating effect on the relationship between SMAP and Business performance?

In order to address these questions, we first developed a conceptual model to analyze the relationship of SMA practices (SMAP), customer engagement and business performance and then tested the model empirically through Structural Equation Modelling (SEM) using a sample of 281 responses.

The remainder of this paper is organised as follows: Sections 2 and 3 are concerned with a review of the literature, research framework and hypothesis development. In section 4 we present and clarify the research methods. Section 5 presents and discusses the results of the empirical data analysis. Section 6 represents the discussion and implications of our research; while Section 7 presents its main conclusions and outlines avenues for future research.

2. Literature Review

2.1 Social media analytics practices (SMAP)

According to Zeng *et al* (2010), SMA "is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data in order to facilitate conversations and interactions…to extract useful patterns and intelligence…". SMA has evolved to become an important driver for acquiring and spreading information in different domains (*Stieglitz et al*, 2018). SMA monitors and analyses data gathered from blogs, forums, Facebook, Twitter, YouTube, etc. all of which contain very significant and valuable information about consumer perceptions, competitors, products, brands and services necessary in making business decisions (Sivarajah *et al*, 2019; Anjanita, 2017). SMA has the potential to provide real-time feedback and actionable insights to help organisations in their decision-making processes (Umar, 2014). SMA has also been defined as a collection of tools, systems, and/or frameworks that facilitate the above types of activity (Grubmüller, Götsch, *et al*. 2013 and Grubmüller, Krieger, *et al* (2013). Yang, *et al* (2011) states that SMA is concerned with developing and evaluating informatics tools and frameworks to measure the activities within social media networks from around the web. Kurniawati *et al* (2013) note the following benefits based on a review of 40 SMA "success stories" from companies such as IBM, SAS, and SAP: 1) Improved marketing strategies (75% of the cases), 2) Better customer engagement (65%), 3) Better customer service (35%), 4) Better reputation management and brand awareness (30%), 5) Product innovation (30%), 6) Business process improvement (25%) and 7) Discerning new business opportunities (20%).

Moreover, Bekmamedova and Shank (2014) note that successful use of social media analytics practices depends on three key organisational practices: 1) Customer management

2) Performance management and 3) Process management. Knowing your customer is one of the most fundamental rules of the retail business and SMA provides retailers with a wealth of information about their customers (Sivarajah *et al*, 2019; Anjanita, 2017). Retailers can better understand customer behaviour by combining intelligence acquired by social media platforms with traditional customer intelligence (Sigala & Chalkiti, 2015). Customer management can be seen primarily in terms of an organisation's capability to understand its customer expectations and market intelligence (Ray *et al*, 2005).

To gain competitive advantage retailers need to monitor and analyse customer-generated content on various social media sites (Lee, 2018). This is important in a SMA context as it enables organisations to improve their customer and market intelligence. Mithas *et al* (2011) define performance management as "… an organization's capability to design and manage effective performance measurement and monitoring systems to support the communication and decision of performance to appropriate stakeholders". If SMA insight is exercised effectively, the organisation will be able to measure its business impact and execute relevant competitive actions (Bekmamedova & Shanks, 2014). Mithas *et al* (2011) further define process management as "…an organization's capability to achieve speed, flexibility and frugality through an effective design and managing the key processes". In the SMA context process management practices ensure that SMA and insights are integrated with the appropriate business processes and relevant metrics to develop and control it (Bekmamedova & Shanks, 2014). SMA has the potential to help businesses understand their audience using social data. SMA tools are also helpful in surfing most social media channels and social networks (Quantzig, 2019).

Customer engagement is extremely important to business performance because without the nature of close dyadic buyer-seller relationship the organisation will not be in a position to gauge the success or otherwise of its business model. By engaging with its customers on a close, continuous, and interactive manner the organisation will not only acquires "real time" information about what is happening in the marketplace, it enables it to have the appropriate data that it then uses as input to its overall strategic development process. There appear therefore to be a clear and significant relationship between customer engagement and business performance. SMA is a powerful tool that enables the organisation to have a better understanding of the significance of the relationship between customer engagement and its own performance in the marketplace.

2.2 Customer Engagement (CE)

CE enables organisations to interact, participate and influence the conversation around their brand. Additionally, effective CE strengthens brand loyalty and influences the discussion and purchase behaviour of the customer (Carr, 2017). Users who have a high appreciation of big brands engage with these brands through “liking”, “sharing” and commenting on them on social media (Araujo and Neijens, 2012; Lin and Lu, 2011, Ruiz-Mafe et al., 2014). CE is built and rebuilt with every brand interaction, whether it relates to making a purchase, reading a social media post or any exposure to the brand (Jacob and Madhav, 2018). Cambra-Fierro Jesús *et al* (2015) define CE as a set of “customer behaviors vis-à-vis the firm – both transactional (loyalty, repurchase intention) and non-transactional (commitment, word-of-mouth, referrals, blogging, etc.) in nature – which guarantee future sales volumes, generate positive publicity and bolster brand reputation”. Understanding the nature of customer/brand relationship has become much more complex with the introduction of social and video sharing platforms such as Facebook, Twitter, blogs, YouTube and Vimeo (Brodie *et al*, 2013; Hollebeek *et al*, 2014). Hollebeek (2011b) defines ‘customer brand engagement’ as “…the level of a customer’s motivational, brand-related and context-dependent state of mind characterised by specific levels of cognitive, emotional and behavioural activity in brand interactions…” (Hollebeek, 2011a: 24). CE is mainly focused on interaction and participation of customers (Wagner and Majchrzak, 2007; Nambisan, 2002). Mollen and Wilson (2010:152) define customer brand engagement as: “a cognitive and affective commitment to an active relationship with the brand as personified by the website or other computer-mediated entities design to communicate brand value”. Brian Haven of Forrester Research (2008) has articulated the definition for engagement as the level of involvement, interaction, intimacy and influence one individual has with a brand over time. “Engagement goes beyond reach and frequency to measure people’s real feelings” (Forrester Research, 2008). Involvement – i.e. the primary point and reflects pragmatic involvement between a person and level of interest in a brand. Thomson *et al*, (2005: 271) define involvement as “a state of mental readiness that typically influences the allocation of cognitive resources for a consumption object, object, or decision”. Involvement can be defined as “perceived relevance of the object based on inherent needs, values, and interests” (Zaichkowsky 1985:342). De Valck *et al* (2009) emphasise the Internet’s capability to act as a medium that enables customers to access online content to communicate with companies. When discussing the introduction of online communities, the concepts of participation and interaction are the most-used ones. The significant contribution people make through comments on company blogs, requests for product information, social media nexus and discussions in forums are different interaction activities. Dholakia *et al* (2004) suggest that 'participation' in an on-line community should be seen as a product of the frequency and duration of community visits and that this is moreover similar to the definition of interaction proposed by Hollebeek (2011a) and Kuo and Feng (2013). CE is extremely important to business performance because without the nature of close dyadic buyer-seller relationship the organisation will not be in a position to gauge the success or otherwise of its business model. By engaging with its customers on a close, continuous, and interactive manner the organisation will not only acquires "real time" information about what is happening in the marketplace, it enables it to have the appropriate data that it then uses as input to its overall strategic development process.

2.3 Business performance (BP)

Many definitions of what constitutes a BP can be discerned from the literature (see, for example, Alchian & Dernsetz, 1972; Flapper, Fortuin and Stoop; 1996; Daft, 2000; Al- Marri et al, 2007; Jing and Avery, 2008). However, in its crystallised form BP has come to be regarded primarily in terms of how the organisation is meeting its objectives seen from the perspective of how it creates value and disseminates that value to its own customers in an optimal manner. A number of studies have attempted to measure BP using financial returns (return on investment, for example) or market-related criteria such as increase in market share, overall competitive position of the organisation in the marketplace, and so on (Stock *et al*, 2000; Chen and Poulaj, 2004; Flynn *et al*, 2010). The definition which we have retained and utilised in the present research is that which relates to value creation, enhancement, and dissemination to customers which, in turn, leads to BP through CE.

3. Research framework and hypotheses development

The conceptual framework for this study is primarily derived from the analysis of many success stories published by SMA vendors and academic resources (SAS, 2011; Cognizant, 2014; IBM, 2013; SAP, 2014; Traphagen, 2015; York, 2017). Moreover, from our own discussions with industry experts and academics we postulate that a possible strong relationship exists between SMAP, CE and BP. For this purpose, SMAP are expected to have a positive and direct relationship with CE. It is also assumed that there is a positive relationship between SMAP and BP mediated by CE. Moreover, it is believed that SMAP may also have a direct and positive relationship with BP.

As part of the preliminary study and to understand the nature, dimensions, scope and items of SMAP, CE and BP, we identified experts from the retail and IT Industries as well as academia that were closely associated with social media and CE-related initiatives in the retail and IT industries across India. We invited a focused group of 10 persons comprising four business executives, two customer relationship management (CRM) experts, two analytics professionals and two marketing professors. These experts had a vast amount of information and knowledge on the benefits of implementing SMAP in the retail and IT industries. Based on extensive review of the literature, coupled with analysis of online comments and combining the initial recommendation of field and focus group interviews, we defined the dimensions and list of items which are considered relevant for inclusion in the data collection instrument (closed-ended structured questionnaire) in order to measure SMAP, CE and BP (for details please refer to the Appendix of the paper).

SMAP is conceptualised as a three dimensional construct which are: customer management, process management, and performance management. CE is conceptualised as a four dimensional construct which are: involvement, interaction, intimacy, and influence; and BP is conceptualised as a two dimensional construct – i.e. financial performance and market performance. Subsequently, the dimensions and items of SMAP, CE and BP were verified using confirmatory factor analysis and validated on different data sets during our pilot study phase. According to Churchill (1979), it is very important to identify theoretical relationships between any newly proposed construct and other conceptually related, but distinct concepts. In the following section, we find constructs related to SMAP that lead to the formulation of the study's hypotheses.

The traditional way of measuring and managing CE has been to conduct customer surveys. This metrics is complex and takes a long time to achieve. With Social Data Analytics retailers can analyse how engaged customers are and how efficient customer relationship management (CRM) is by counting how many of their customers "like" their publications, "share" them, comment on them or speak about a brand. Retailers get a better understanding of both their existing and prospective customers and embrace a new way of engaging with them (Anjanita, 2017). The impact of social media extends across all industries and is particularly prevalent in the retail industry where it has completely changed the brand-customer relationship into customer-centric interaction. Developments in SMA suggest that it is important for retailers to hold discussions with their customers, improve their facilities, radar in on content, and bankroll on engagement insights. SMA is increasingly being used to generate deep customer insights based on buying patterns, demographics, web behaviour, social media, and product affinities. According to TCS (2014) this enables tracking of online mobile engagement and conversations on social media in order to improve customer retention and increase vista conversion. It also improves the accuracy of demand vaticinator, thus helping to better anticipate customer needs. The opportunities for delivering this type of engagement have never been greater. Customers have accumulated a wealth of information that businesses can analyse in order to create data-enriched insights into customer behaviour and need. The amount and context of data ultimately drive more personal nexus between the customer and business, enabling even deeper and more authentic engagements (SAP, 2014). However, there is currently no empirical evidence in the literature which sufficiently demonstrate that SMA will indeed improve CE.

Based on the preceding discussion we have developed the following hypotheses for our research:

H1. There is a positive relationship between social media analytics practices and customer engagement in the Indian retail industry.

CE recognises that organisational buying from consumers is important that they will influence BP (Carr, 2017). Employee engagement leads to CE and that CE, in turn, can lead to performance increases of up to 240% (Stephenson, 2014). A study conducted by People Metrics on 10000 customers in 2008 (Peoplemetrics, 2008) reveal that organisations with highly engaged customers yielded 8 % return on investment (ROI) above the industry average; while organisations with less engaged customers experienced a decrease in their profit margins by as much as 23% below the industry average. Highly engaged customers led their organisations to grow 13% above the industry average and low engaged customers led to organisations declining profit margins by 36% as against the industry average (Peoplemetrics, 2008). Moreover, it was concluded that retail organisations that engaged with their customers had better financial results than those who failed to engage with their customers (People Metrics, 2010). Engagement takes many forms, including content consumption, website page views, email opens, paid and organic search clicks, call centre interactions, “likes” and “follows”, tweets and re-tweets, and referrals (McLeester, 2014). CE is a one-dimensional concept and as such, focuses on either the emotional, the cognitive, or behavioural aspects of engagement. According to Brodie *et al* (2011) the behavioural dimension in particular appears dominant within the one-dimensional perspective. Engagement behaviourally summarises the impact of marketing/branding communications activities in the hearts and minds of consumers in a manner that leads to higher sales margins, market share, market value, and cash flow (ARF, 2006). According to the Economist Intelligence Unit (2007), more CE translates into improved customer loyalty (80%), increased revenue (76%) and increased profits (75%). However, there is no empirical evidence available in the literature that sufficiently demonstrates that CE will improve BP. As a result:

H2. There is a positive relationship between customer engagement and business performance in the Indian retail and IT industries.

The use and importance of social media by businesses is expected to flourish in the coming years; and in particular the use of analytical capabilities to analyse and interpret vast amounts of online information to gain customer and business documents will assume a heightened importance (HBR, 2010; IBM, 2013). KIA motors and The Royal Bank of Canada have achieved product innovation, customer service improvement, and identification of new business opportunities through SMA (Kite, 2011). SMA is moreover about combining the social data and transforming it into meaningful information which can then be used in an organisation to enrich brand visibility and improve top line sales figures. According to Shiauet *et al* (2017), “……Contemporary firms should pay additional attention to social network users and effectively utilize social networks to enhance firm performance….” However, there is currently no empirical evidence available in the literature which demonstrates that SMA practices will improve BP. As a result:

H3. There is a positive relationship between social media analytics practices and business performance in the Indian retail and IT industries.

Considering hypothesis H1-H3, the possibility also existed that customer engagement may serve as a mediating role between social media analytics practices and business performance.

A number of studies have examined how SMAP can generate more profit and sales revenue by engaging customers (Parveen et al., 2015, Kamboj et al., 2016, Setyani et. al., 2019, SAS, 2011; Cognizant, 2014; IBM, 2013; SAP, 2014; Traphagen, 2015; York, 2017) but considering this central role of CE we do not find empirical evidence in literature studies that sufficiently demonstrates that SMA will improve BP through CE. Accordingly, it was finally hypothesised that:

H4. Customer engagement has a mediating role on the relationship between social media analytics practices and business performance.

Figure 1 presents the conceptual model used in the present research. The model proposes that SMAP have a positive relationship with BP both directly and indirectly through CE.

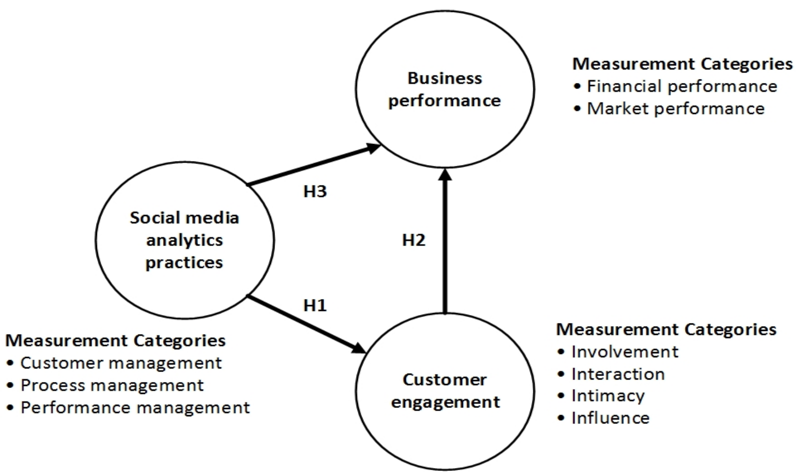


Figure 1: A conceptual model of research and hypotheses development

4. Research Methods

4.1 Instrument development

Our questionnaire uses a scale to measure various constructs of the research model depicted in Figure 1. Subsequent to literature review and expert opinion, an initial draft of a structured closed-ended questionnaire was developed which contained 33 items (questions). These were divided as follows: 15 items related to SMAP, 12 items for CE, and 6 items for BP. A five-point Likert scale was used to specify the respondents' level of agreement to the statements. Items of SMAP were evaluated on a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5); items of CE ranged from Never (1) to all the time (5); and items of BP ranged from significant decrease (1) to significant increase (5). Details of scale items are indicated in the Appendix to this paper. Furthermore, reliability and validity of the questionnaire using two pilot tests were carried out. Feedback was gathered after each pilot test and the questionnaire was refined in response to the feedback. During the first pilot test the 33 item questionnaire was sent to the 7 subject-matter experts in order to gauge their reaction on the wording and content of the items. In the second pilot test the questionnaire was sent to 65 Management consultants and Digital marketing executives. Data received from the questionnaire were statistically processed. Cronbach alpha and Confirmatory Factor Analysis (CFA) was conducted to check reliability, validity, and statistical fit as well as finalise the list of measures.

4.2 Data Collection

Initially, 400 respondents were contacted by telephone to ascertain whether or not their organisation had implemented SMAP if so whether it would be interested in participating in our research, the right person within the organisation we could contact, etc. As per the response of the organisation, Digital Marketing Executives, Digital Marketing Specialist, Management consultants, Analytics Managers, CRM Managers, Marketing Director, engagement Managers who had responsibility for developing a digital marketing strategy for the retail organisation were contacted and asked to provide responses to the questionnaire. We used a mixed-mode survey procedure which was adapted from Dillman (1978, 2007). The questionnaire was sent to the target population through mail directly to the respondents, through personal phone interviews, or through Google platforms.

We received a total of 337 responses of which 56 were rejected (for various reasons) leaving a total of 281 usable responses which constitute the data for our empirical analysis.

4.3 Sample profile

Of the 281 respondents, 33.1% were females, while 66.9% were males. Around 52 % respondents were between 27 and 35 years of age and 48% were between 35 and 50 years old. More than half of the respondents (54%) had MBA degree and 39% of them held B.

Tech./B.E. qualifications. Of the total number of respondents 42 per cent were Digital Marketing Executives/Digital Marketing Specialists, engagement Managers/Marketing Directors; 40 percent were Management consultants, Analytics Managers, CRM Managers; and the remainder occupied intermediate positions in the retail and IT industries. It was observed from the final sample that 95% of the respondents had knowledge of SMA. It can be concluded from the profile of the respondents that they were either involved in retail marketing strategy or implementation of SMA.

4.4 Item generation

In order to test the hypothesised relationship between the constructs we needed to generate item pool depicted in the theoretical model (Figure 1) – i.e.: 1. SMA, 2. CE, 3. BP. The purpose of generating the items pool was to achieve content validity of the constructs by reviewing the literature and consulting with experts on the subject. This is consistent with the postulation of Churchill (1979) which suggests that measurement items for a scale should cover the content domain of a construct.

To generate the items for SMA we have reviewed previous related research (Bekmamedova and Shank, 2014; Zeng *et al*, 2010; Zeng *et al*, 2010; Grubmülleret *et al*, 2013; Yang *et al*, 2011; Kurniawati *et al*, 2013; Sigala & Chalkiti, 2015; Lee, 2018; Mithas *et al*, 2011). The research is a rich pool of illustrations, definitions and items of SMAP. As indicated earlier a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5) was used with regard to experts' perception of SMAP in the organisation. Experts were instructed to retain items based on clarity of words and relevance related to SMAP. Items that were not relevant and not clear were deleted from the items pool. All of the respondents were subject-matter experts in the domain of digital marketing. Items were created in 3 groups as per the corresponding three sub-dimension proposed in the theoretical model (Figure 1).

To generate the items for customer engagement we have reviewed previous related research (Brodie *et al*, 2013; Hollebeek *et al*, 2014; Hollebeek, 2011a; Hollebeek, 2011b; Wagner and Majchrzak, 2007; Nambisan 2002; Mollen and Wilson, 2010; Haven of Forrester Research, 2008; Haven Brian, 2007, Astute, 2019). A 5-point Likert scale ranging from Never (1) to all the time (5) was used in reference to experts' perception in order to track customer engagement in their organisation. Items were created in 4 groups as per corresponding four sub-dimension proposed in the theoretical model (Figure 1).

To generate the items for business performance we have reviewed previous related research (Venkatraman & Ramanujan, 1987; Al-Marri *et al*, 2007; Combs *et al*, 2005; Jing & Avery, 2008; March and Sutton, 1997; Stock *et al*, 2000; Flynn et al., 2010).). A 5-point Likert scale ranging from Never (1) to all the time (5) was used in reference to experts' perception in order to measure business performance in the organisation. Items were created in 2 groups as per corresponding two sub-dimension proposed in the theoretical model (Figure 1).

5. Data analysis and results

We performed Structural Equation Modelling (SEM) using AMOS 18.0 to analyse the empirical data. SEM is a statistical method for analysing causal relationships in a set of constructs represented by multiple measurable variables/items in a single model. SEM helps in analysing theoretical relationship between different constructs. It comprises two components, namely: the measurement model and structural equation model (Blunch, 2008). The measurement model defines how latent variables are measured or operationalised using the observed variables; it provides the validity and reliability of measures used in representing the latent variable. The structural equation model on the other hand explains the assumed causation in the set dependent and independent constructs developed from the conceptual model (Gefen *et al*, 2000; Hair *et al*, 2010).

5.1 Assessment of the First-order measurement model

The conceptual model used in this study contains multidimensional constructs. First, we perform confirmatory factor analysis (CFA) for each construct to assess the validity and reliability of the first-order measurement model. Before hypothesis testing it is mandatory to test the first-order measurement model for validity and reliability (Fornell & Larcker, 1981). Figure 2 shows a first order measurement model which focuses on the relationship between dimensions/sub-construct and items.

The factor loadings of latent to observed variables should be above 0.50 (Hair *et al*, 2010). Three items (i.e. PEM4 = 0.48, PEM7= 0.35 and INF1 =0.32) were deleted from the model because of low factor loading (<0.50). After deleting the 3 items from first-order measurement model, all measures were analysed for reliability and validity. The reliability of these constructs was evaluated using Cronbach's coefficient alpha and the value should be above 0.7, indicating a reliable measurement instrument (Nunnally, 1978). To assess the construct validity Churchill (1979) suggests that the convergent and discriminant validities should be examined. Therefore, we measure convergent validity by composite reliability and average variance extracted measures. Composite reliability is a measure of the internal consistency of the construct in a scale; while average variance extracted can be defined as the extent of the variance of variable which is explained by the latent constructs. Suggested value of CR should be greater than 0.7 (Hair, 2010) and AVE should be greater than 0.5 (Fornell and Larcker, 1981).

Results of Table 1 clearly demonstrate the adequate reliability and convergent and discriminant validity of all the sub-constructs.

Discriminant validity is the degree to which variables in different constructs are different from each other. It means that variables in different constructs have low correlation between themselves. According to Fornell & Larcker (1981) in order to establish discriminant validity, "the square root of a construct's AVE must be larger than the inter-construct correlations". Table 2 explains the results of discriminant validity. The element in diagonal represents the square root of the average variance extracted. All sub-constructs showed more variance with their indicators than with other sub-constructs. The square root of AVE exceeds the correlation between other constructs. These results imply satisfactory discriminant validity. After testing the measurement model with all the parameters mentioned above we can confirm that the model is reliable and valid.

Table 1: Reliability and Items loading						
Dimensions/ Sub-construct	**Items**	**Standard factor loading**	**Cronbach (α)**	**Composite reliability**	**Average variance extracted**	**Average shared variance**
Customer Management (CM)	CM1	.725	0.800	0.801	0.505	0.095
	CM2	.689				
	CM3	.749				
	CM4	.670				
Process management (PM)	PM1	.703	0.806	0.809	0.515	0.099
	PM2	.671				
	PM3	.751				
	PM4	.743				
Performance management (PEM)	PER1	.739	0.825	.829	.493	0.093
	PER2	.754				
	PER3	.739				
	PER5	.662				
	PER6	.740				
Involvement (IN)	IN1	.787	0.815	0.815	0.595	0.210
	IN2	.814				
	IN3	.708				
Interaction (INT)	INT1	.849	0.870	0.873	0.697	0.216
	INT2	.896				
	INT3	.752				
Intimacy (INM)	INM1	.729	0.870	0.862	0.677	0.228
	INM2	.894				
	INM3	.836				
Influence (INF)	INF2	.778	0.805	0.807	0.676	0.212
	INF3	.865				
Financial performance (FP)	FP1	.792	0.743	0.744	0.594	0.092
	FP2	.880				
Market Performance (MP)	MP1	.721	0.884	0.888	0.666	0.102
	MP2	.819				
	MP3	.867				
	MP4	.850				

Table 2: Discriminant validity of the first-order measurement model									
	CM	**PM**	**PER**	**IN**	**INT**	**INM**	**INF**	**FP**	**MP**
CM	**0.709**								
PM	0.607	**0.718**							
PER	0.539	0.568	**0.702**						
IN	0.042	0.132	0.042	**0.771**					
INT	0.103	0.195	0.264	0.706	**0.835**				
INM	0.017	0.114	0.062	0.737	0.755	**0.823**			
INF	0.068	0.176	0.085	0.756	0.676	0.784	**0.822**		
FP	0.215	0.012	0.093	0.146	0.144	0.184	0.069	**.770**	
MP	.184	.014	.201	.166	.249	.206	.088	.760	**.816**

5.2 Assessment of the Second-order measurement model

SMAP, CE and BP was conceptualised as a second-order model composed of 3, 4 and 2 dimensions respectively.

Second-order models are potentially applicable when (a) the lower order factors are substantially correlated with each other, and (b) there is a higher order factor that is hypothesised to account for the relations among the lower order factors. The reliability and validity of the second-order model can be measured similar to first-order model.

The results in table 3 confirm high reliability of the second-order measurement high validity in terms of convergent and discriminant validity. Therefore, we conclude that the second-order measurement model is internally consistent and reliable as suggested by Fomell and Larcker (1981).

Table 3: Reliability and items loading					
Construct	Sub-construct	Loading	Cronbach (α)	Composite reliability	Average variance extracted
Social Media Analytics Practices (SMAP)	Customer Management (CM)	0.760	0.871	0.804	0.591
	Process management (PM)	0.786			
	Performance management (PEM)	0.726			
Customer engagement (CE)	Involvement (IN)	0.847	0.921	0.918	0.737
	Interaction (INT)	0.829			
	Intimacy (INM)	0.901			
	Influence (INF)	0.856			
Business performance (BP)	Financial performance (FP)	0.785	0.885	0.886	0.798
	Market Performance (MP)	0.988			

Similar to the first-order measurement model, we tested the discriminant validity for the second-order model. Table 4 shows that all of the diagonal values exceed the squared inter-construct correlations. Therefore, we conclude that the first-order construct can be explained by the second-order construct.

Table 4: Discriminant validity of the second-order measurement model			
	CE	**SMAP**	**BP**
CE	**0.859**		
SMAP	0.163	**0.769**	
BP	0.213	0.069	**0.893**

In order to analyse the statistical fitness of the structural model, many fitness indices like the comparative Fit Index (CFI), the goodness-of-fit index (GFI), Normed fit index (NFI), Tucker-Lewis Index (TLI) and Root Mean Square of Error Approximation (RMSEA) are used. The ideal values indices of X2/df should be less than 3, CFI,GFI, NFI, and TLI should be more than 0.9 and the RMSEA value must be lower than 0.08 (Geffen, 2000). Table 5 represents a brief summary of goodness-of-fit indices of second-order measurement model. The respective value of X2/df, CFI, GFI, NFI, and TLI are 1.498, 0.944, 0.863, 0.851and 0.938. The value of RMSEA is 0.046. Although the GFI and NFI value of 0.876 and .860 could not meet the criteria, the values are the closet threshold thus representing an acceptable model fit.

Table 5: Summary of goodness-of-fit Indices for Measurement Model						
Model Fit Index	Chi-square/ Degree of freedom	CFI	GFI	NFI	TLI	RMSEA
Model	1.498	0.944	0.836	0.851	0.938	0.046

Based on the preceding it can be concluded that the second-order measurement model represents a good fit and as such we can proceed to testing the structural model using SEM.

5.3 Assessment of the Structural Model

A structural model was developed using SEM to examine the hypothesised conceptual research model (Figure 2). A brief summary of fitness indices for the structural model are shown in Table 6. The values of X2/DF, CFI, GFI, NFI, and TLI are 1.498, 0.944, 0.963, 0.851 and 0.938 respectively. The RMSEA shows a value of 0.046. As per fitness indices suggested by Gefen (2000), GFI and NFI could not meet the criteria but the values were very close to the threshold. Thus we can conclude that the structural model is accepted as per fit indices. This allows us to continue examining the research hypothesis defined in our model.

Table 6: Summary of Goodness-of-Fit Indices for Full Model						
Model Fit Index	Chi-square/ Degree of freedom	CFI	GFI	NFI	TLI	RMSEA
Model	1.498	0.944	0.863	0.851	0.938	0.046

Table 7 summarises the properties of the structural model (standardised path coefficients (β) and hypotheses result). The level of significance (α) is set at 0.05.

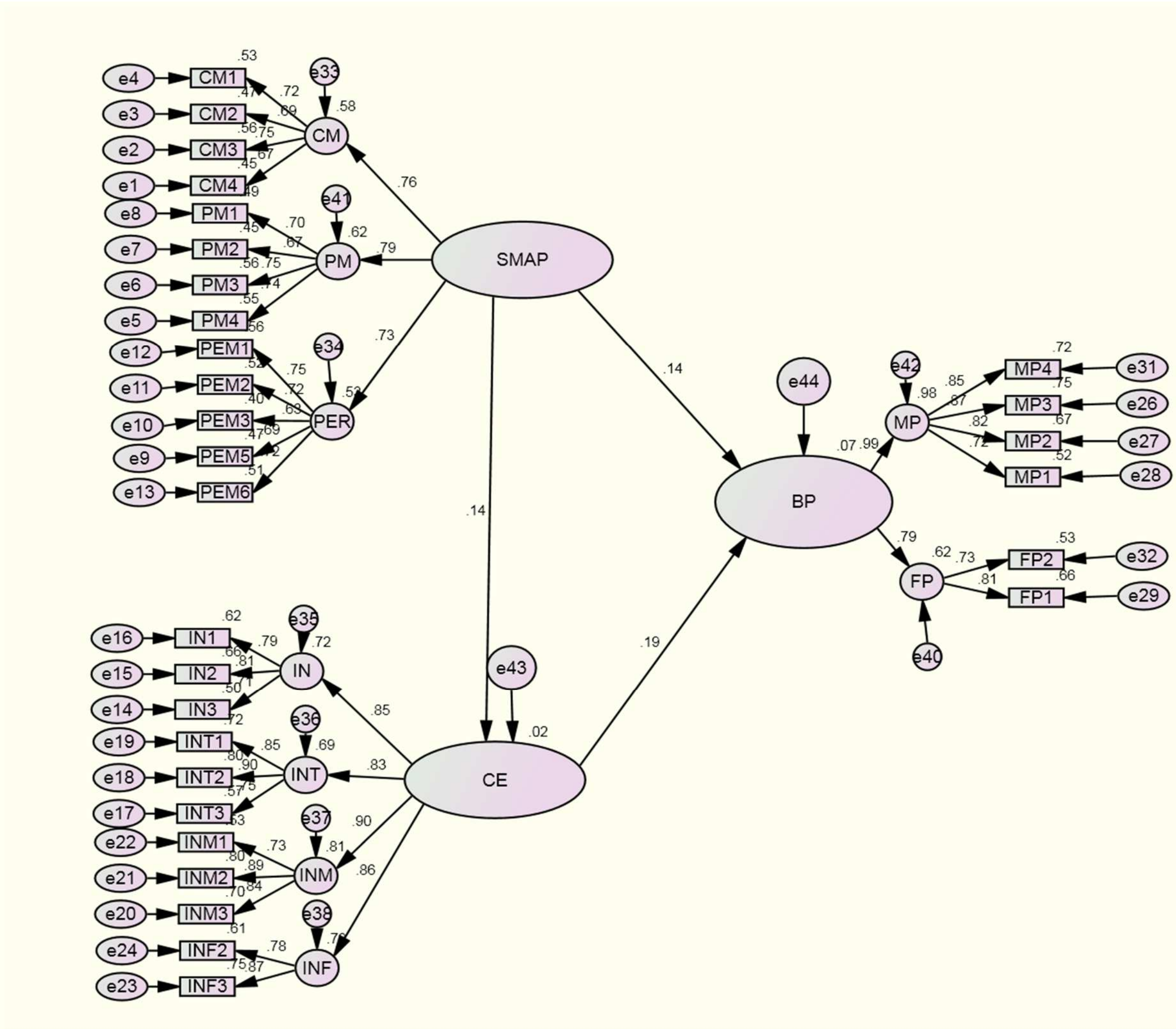


Figure 2: Structural model

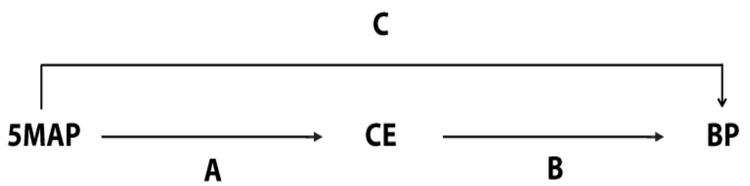
Table 7: Summary of testing of hypotheses						
Hypothesized path			**Estimates(β)**	Unstandardized Regression Weight	P	Result
CE	<---	SMAP	0.140	0.193	***	Supported
BP	<---	SMAP	0.143	0.174	***	Supported
BP	<---	CE	0.194	0.170	***	Supported

Notes: β = *standardised beta coefficients;* P < 0.05, ***p<.005

The results indicate that the strategic use of SMAP is significantly and positively related to CE (β=0.140; $p < 0.05$), providing support for hypothesis 1. There is a positive relationship between CE and BP (β=0.143; $p < 0.05$), demonstrating that hypothesis 2 is also supported. The results indicate that SMAP have a positive relationship to BP (β=0.143; $p < 0.05$), indicating that hypothesis 3 is supported. The estimates are consistent with the expectations due to the fact that the relationship is significant ($p < 0.05$) and in the anticipated direction.

5.4 Testing for mediation

In order to examine the mediation effect on customer engagement (Baron and Kenny, 1986), a four-step regression method was used. In our model SMAP are denoted by SMAP, BP and CE. The diagram below shows the mediation path of SMAP and BP.



Accordingly, the mediation effect in which SMAP leads to BP through CE is called the indirect effect. The indirect effect represents the portion of the relationship between SMAP and BP that is mediated by CE.

Step1: We conduct a regression analysis to identify the relationships between SMAP and BP to test the significance of path C.

$$BP = C_0 + C * SMAP + ?? \qquad \text{------------------------------------}(1)$$

Step 2: We conduct a regression analysis to identify the relationships between SMAP and CE to test the significance of path A

$$CE = A_0 + A1 * SMAP + ?? \qquad \text{------------------------------------(2)}$$

Step 3: We conduct a regression analysis to identify the relationships between CE and BP to test the significance of path B

$$BP = B_0 + B1 * CE + ?? \qquad \text{------------------------------------(3)}$$

Step 4.

$$BP = \propto + C' * SMAP + B * CE + ?? \qquad \text{------------------------------------(4)}$$

According to Baron and Kenny (1986), a full mediation occurs if the effect of a mediating variable (CE in this context) remains significant after controlling for independent variable (SMAP in this context). On the other hand, a partial mediation is deemed to have occurred if the relationship between the independent variable and the dependent variable is still significant after controlling for the effects of the intervening variable (i.e. SMAP construct and CE significantly predict BP). The summary of the mediation effect is given in Table 4. The results in Table 4 show that CE partially mediates the relationship between SMAP and BP (Direct effect 0.143***; Indirect effect 0.027***, supporting hypothesis 4). The mediation results indicate that CE plays a significant role in boosting the effective use of SMP on BP.

Table 8: Mediation effect of Customer Engagement

Hypothesis path	Direct effect	Indirect effect	Result
SMAP -> CE->BP	0.143***	0.027***	Partial mediation

6. Discussion, implication and future research

The objective of this research has been to investigate and produce knowledge about the nature of causality existing between SMAP, CE and BP. The research has empirically generated valuable findings and has established that there is a causal relationship between SMAP, CE and BP consistent with Anjanita (2017) and Stephenson (2014) postulations. Our findings also confirm the mediation effect of CE on the relationship between SMAP and BP consistent with Parveen *et al* postulations (2015). In the current highly competitive global marketplace CE has significant imperative for businesses. With SMAP businesses can easily analyse the on-going interactions between an organisation and its customers. If the customer is engaged (or interested) in what the organisation or brand is doing it is clear that the customer will be more likely to engage with the business in term of purchases and contribute to improving its performance in the marketplace. CE is generally defined by the number of different actions on the part of the customer including purchases, social sharing, and referrals. SMAP analyses the nature and extent of the relationship that exists between an organisation and its customers which, in turn, is critical for a more precise analysis of the organisation's on-going growth and success. The results of this research lead to a number of important findings which have both theoretical and managerial implications.

6.1 Theoretical implications

Social media analytics is still a relatively new research area. The present research has provided a theoretical model that identifies positive and significant relationship between SMAP, CE and BP. Its main contribution is supported by the conceptual model depicted in Figure 1. The conceptual model provides a foundation for future research in the area of SMA. In addition, we reinforce the findings of previous studies and several "white papers" which have pointed out the relationship between SMAP, CE and BP (SAS, 2011; The Enterprise Strategy Group, 2013; Cognizant, 2014). However, to the best of our knowledge this relationship has not yet been empirically investigated within the context of the Indian retail and IT industries. As such the present research can be said to have made a contribution in enhancing our overall understanding of these matters within the context of these industries in India. Furthermore, this research contributes to the existing body of knowledge by analysing the mediation effect of CE on the relationship between SMAP and BP.

6.2 Managerial implications

The objective of this research has been to examine the nature of relationship (if any) between SMAP, CE and BP. This research has empirically generated valuable findings and has established causal relationship between SMAP, CE, and BP. These results have enabled the validation of the hypothesis that strategic use of SMAP has a positive relationship between CE and BP. SMA helps the retail and IT organisations magnify their business presence, run a smart social media campaign; reduce customer services/support costs by using social media monitoring process. It also enables the building of an active community by administering an online conversation and taking customer responses and opinions on a particular product and its services. In today's highly competitive global marketplace, businesses want to discover their customer's likes, dislikes, preferences and habits. Moreover, they want to track an integrated picture of customers across many contact points in the marketplace like, for example, how many leads can be brought in through social media efforts in addition to monitoring customer activity and interactions. Apart from examining customer-centric integrated information, businesses also want to know their competitors in order to evaluate the gaps by comparing competitors' strengths and weaknesses. In a nutshell, social media provides a wealth of information to retail and IT organisations. Real time SMA can help the retail and IT industries make effective business decisions. If retail and IT organisations are serious about social media marketing initiatives, then strategic use of SMA can enable these organisations to evaluate and understand customers' interactions, feedback and provide timely and effective responses.

CE has in recent years assumed an added significance for retailers. CE involves regular customer interaction which, needless to add, is an on-going activity. By engaging customers effectively organisations are able to gain competitive advantages, increase customer loyalty, enhance revenue, and manage their operational costs in an optimal manner. Highly engaged customers resolve issues with the organisation directly rather than complaining publicly about unsatisfactory shopping experiences. Moreover, highly engaged customers are more likely to encourage their friends and family to become customers. SMA enables more effective personal interaction, timely response to feedbacks, identifying and addressing the queries as soon as they arise, and so on. It also helps in delivering consistent, contextual, and adapted experiences using an effective data-driven strategy for engaging the customers with the retailers. SMA provides the capability to the retail organisation to understand what kind of media and content is driving customer engagement and how customers react. It further enables the sales and marketing team to develop more efficient and effective customer loyalty programs, product development and enhancement, pricing, and other important sales, marketing, and customer support activities. SMAP is serving retailers better in terms of aligning market strategies with CRM and actionable insights. SMAP measure the overall strategy of an organisation. Therefore, it becomes imperative for retailers to capture consumer data from social media in order to understand attitudes, opinions, and trends and manage online reputation to better serve their customers. Strategic use of SMA can benefit retailers in tracing the quality and quantity of a brand's reference across the entire social media; follow exchanges on chat forums, blogs and other social channels. By engaging customers effectively, retailers can gain competitive advantage, enhance business performance, and reduce their operational costs.

7. Conclusions, Limitations and Future research

The most significant conclusion of this research is that there is a positive relationship between SMAP and BP in which CE plays a key mediation role. It examines four research questions and to address these questions a comprehensive model was developed and tested using Structural Equation Modelling (SEM) analysis. This research provides empirical justifications for the existence of a causal relationship between SMAP, CE, and BP. It provides empirical evidence to support the theoretical and prescriptive statements in the literature. Its major contribution, however, is that it demonstrates, from an empirical perspective, the importance of SMAP for the retail and IT industries.

Despite these encouraging empirical findings, this research can also be said to have a number of limitations which affect its generalisation. Firstly, although we have considered widely accepted items of SMAP, CE and BP derived from the literature there is the possibility that we may not have included in the research some of the items which are less common in the literature. Moreover, the findings of our research relate solely to Indian retail and IT industries. As such it may not constitute a sufficient basis for generalisation. These limitations, however, pave the way for future research. To enhance the generalisation of the findings, the model used in this research can be tested by conducting cross-country studies in many geographic regions.

References

Al-Marri, K., Moneim M. Baheeg Ahmed, A., &Zairi, M. (2007). Excellence in service: an empirical study of the UAE banking sector. *International Journal of Quality & Reliability Management*, *24*(2), 164-176

York, A.(2017). What Is Social Media Engagement & Why Should I Care?. Retrieved from https://sproutsocial.com/insights/what-is-social-media-engagement/

Alchian, A. A., &Demsetz, H. (1972).Production, information costs, and economic organization. The *American Economic Review*, 62(5), 777-795.

Anjanita(2017). How to leverage Social Media Analytics for your business?. Retrieved fromhttps://www.analyticsvidhya.com/blog/2017/02/social-media-analytics-business/

Araujo, T. and Neijens, P. (2012). Friend me: which factors influence top global brands participation in social network sites . *Internet Research, 22*(5), 626-640.

Astute. (2019). *What You Need to Know about Measuring Customer Engagement*. Retrieved from https://www.astutesolutions.com/blog/articles/what-you-need-to-know-about-measuring-customer-engagement

ARF – Advertising Research Foundation (2006). Engagement: Definitions and Anatomy. ARF White Paper, J. Plummer (Ed.), 21 March

Bekmamedova, N., & Shanks, G. (2014). Social media analytics and business value: A theoretical framework and case study. In 2014 47thIEEE Hawaii international conference on system sciences, 3728-3737.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182

Blunch, N.J. (2008), “Introduction to Structural Equation Modeling using SPSS and AMOS”, CA: Sage Publications

rodie, R. J., Hollebeek, L. D., Jurić, B., &Ilić, A. (2011). Customer engagement: Conceptual domain, fundamental propositions, and implications for research. *Journal of Service Research*, *14*(3), 252-271.

Brodie, R. J., Ilic, A., Juric, B., &Hollebeek, L. (2013). Consumer engagement in a virtual brand community: An exploratory analysis. *Journal of Business Research*, *66*(1), 105-114.

Carr, P. (2017). Whole enterprise social media for business performance. In *Phantom Ex Machina* (pp. 25-35). Springer, Cham.

Cambra-Fierro, J. J., Melero-Polo, I., & Vázquez-Carrasco, R. (2013). Customer engagement: Innovation in non-technical marketing processes. *Innovation*, *15*(3), 326-336.

Churchill Jr, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 64-73.

Cognizant (2014). “How Banks Can Use Social Media Analytics to Drive Business Advantage,
Retrieved from http://www.cognizant.com/InsightsWhitepapers/How-Banks-Can-Use-Social-Media-Analytics-To-Drive-Business-Advantage.pdf

Combs, J. G., Russell Crook, T., & Shook, C. L. (2005).The dimensionality of organizational performance and its implications for strategic management research. In *Research methodology in strategy and management* (pp. 259-286). Emerald Group Publishing Limited.
Daft, R. L. (2000). *Organization Theory and Design* (7th ed.). South-Western College Publishing, Thomson Learning: U.S.A.
De Valck, K., Van Bruggen, G. H., & Wierenga, B. (2009). Virtual communities: A marketing perspective. *Decision Support Systems*, *47*(3), 185-203.
Dholakia, U. M., Bagozzi, R. P., &Pearo, L. K. (2004).A social influence model of consumer participation in network-and small-group-based virtual communities. *International Journal of Research in Marketing*, *21*(3), 241-263.
Dillman, D. A. (1978).*Mail and telephone surveys: The total design method* (Vol. 19). New York: Wiley.
Dillman, D. A. (2007). Mail and internet surveys: the tailored design method. Hoboken, NJ: John & Wiley Sons.
Dwivedi, Y. K., Kapoor, K. K., & Chen, H. (2015).Social media marketing and advertising. *The Marketing Review*, 15(3), 289-309.
Economist Intelligence Unit(2007), Beyond loyalty Meeting the challenge of customer engagement, Retrieved from http://graphics.eiu.com/files/ad_pdfs/eiu_AdobeEngagementPt_I_wp.pdf
Flapper, S. D., Fortuin, L., & Stoop, P. P. (1996).Towards consistent performance management systems. *International Journal of Operations Management*, 17(7), 27-37.
Flynn, B. B., Huo, B., & Zhao, X. (2010). The impact of supply chain integration on performance: A contingency and configuration approach. *Journal of Operations Management*, *28*(1), 58-71.
Forrester (2008), "Four Steps to Making Engagement Measurement a Reality", retrieved from https://www.forrester.com/Measuring+Engagement/fulltext/-/E-res44421
Fornell, C., & Larcker, D. F. (1981).Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.
Gefen, D., Straub, D., & Boudreau, M. C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems,* 4(1), 1-70.
Grubmüller, V., Götsch, K., & Krieger, B. (2013). Social media analytics for future oriented policy making. *European Journal of Futures Researc*h, 1(1), 1-9.
Grubmüller, V., Krieger, B., & Götsch, K. (2013). Social Media Analytics for government in the light of legal and ethical challenges. In Conference for E-Democracy and Open Government, Donau-Universität-Krems, Austria, pp. 185-195.
Hair, J., Black, W., Babin, B., & Anderson, R. (2010). Multivariate data analysis 7th edth ed. Prentice Hall, Upper Saddle River, New Jersey
Haven, B. (2007), ''Marketing's New Key Metric: Engagement,'' August 8, (Accessed May 15, 2010), Retrieved from http://snproject.pbworks.com/f/NewMetric_Engagement.pdf
HBR. (2010), "The new conversation: Taking social media from talk to action", Retrieved from https://hbr.org/product/the-new-conversation-taking-social-media-from-talk/an/10815-PDF-ENG
Hollebeek, L. (2011a). Exploring customer brand engagement: definition and themes. *Journal of strategic Marketing*, *19*(7), 555-573.
Hollebeek, L. D. (2011 b). Demystifying customer brand engagement: Exploring the loyalty nexus. *Journal of Marketing Management*, *27*(7-8), 785-807.
Hollebeek, L. D., Glynn, M. S., & Brodie, R. J. (2014). Consumer brand engagement in social media: Conceptualization, scale development and validation. *Journal of Interactive Marketing*, *28*(2), 149-165.

IBEF(2018), E-Commerce retrieved from https://www.ibef.org/download/e-commerce-dec-2018.pdf
Internet lives today(2018), Twitter Usage Statistics available at http://www.internetlivestats.com/twitter-statistics/
IBM (2013), “Social media analytics-making customer insights actionable”, retrieved from http://www.mcrinc.com/Documents/Newsletters/201401_SocialMediaAsMarketingTool.pdf
Jacob, J.K., & Bindal, M. (2018). The Impact of Brand Post Strategy and Social Media Engagement on Customer Loyalty, *International Journal of Management Studies*, 5(3-8), 48-55
Jing, F. F., & Avery, G. C. (2008).Missing Links in Understanding the Relationship between Leadership and Organizational Performance. *International Business & Economics Research Journal, 7*(5), 67-78.
Kamboj, S., Yadav, M., Rahman, Z., & Goyal, P. (2016). Impact of social CRM capabilities on firm performance: Examining the mediating role of co-created customer experience. *International Journal of Information Systems in the Service Sector (IJISSS)*, *8*(4), 1-16.
Kapoor, KK, Tamilmani, K, Rana, NP, Patil, P, Dwivedi, YK and Nerur, S (2018). Advances in Social Media Research: Past, Present and Future, *Information Systems Frontiers*, 20(3), 531–558.
Kite, S. (2011). “Social CRM’s a tough, worthy goal”, *Bank Technology News*, 24(6), 30–31
Kuo, Y., & Feng, L. (2013).Relationships among community interaction characteristics, perceived benefits, community commitment, and oppositional brand loyalty in online brand communities. *International Journal of Information Management*, 33, 948-962.
Kurniawati, K., Shanks, G., and Bekmamedova, N. (2013). "The Business Impact of Social Media Analytics," in Proceedings of the 21th European Conference on Information System, Utrecht, The Netherlands, p. 48-61
Lee, I. (2018). Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, *61*(2), 199-210.
Lin, K. Y., & Lu, H. P. (2011).Intention to continue using Facebook fan pages from the perspective of social capital theory. *Cyberpsychology, Behavior, and Social Networking*, *14*(10), 565-570.
March, J. G., & Sutton, R. I. (1997).Crossroads—organizational performance as a dependent variable. *Organization Science*, *8*(6), 698-706.
Traphagen, M. (2015), Why Engagement DOES Matter As A Social (Media Metric, Available at:https://marketingland.com/engagement-matter-social-media-metric-114497
McLeester, C. (2014). “Take a Holistic View of Measuring Customer Engagement”, retrieved from https://www.dmnews.com/marketing-channels/multi-omnichannel/article/13036904/take-a-holistic-view-of-measuring-customer-engagement
Mithas, S., Ramasubbu, N., & Sambamurthy, V. (2011). How information management capability influences firm performance. *MIS Quarterly*, *35*(1), 237.
Mollen, A., & Wilson, H. (2010). Engagement, telepresence and interactivity in online consumer experience: Reconciling scholastic and managerial perspectives. *Journal of Business Research*, 63(9-10), 919-925
Nambisan, S. (2002). Designing Virtual Customer Environments for New Product Development: Toward a Theory. *Academy of Management Review*, 27 (3), 392-413.
Nunnally, J. C., and Bernstein, I. H. (1994).*Psychometric Theory*. New York: McGraw-Hill, Inc.
Peoplemetrics, INC (2010), “Business Outcomes: Customer Satisfaction & Customer Engagement”, retrieved from http://www.peoplemetrics.com/blog/the-importance-of-customer-satisfaction-and-customer-engagement-in-business-outcomes

PeopleMetrics, Inc(2008), Enhancing the Customer Experience and Engagement in Retail retrieved from http://info.peoplemetrics.com/hs-fs/hub/221727/file-2339617306-pdf/PDFs/CE-Study-Retail.pdf

Parveen, F., Jaafar, N. I., & Ainin, S. (2015). Social media usage and organizational performance: Reflections of Malaysian social media managers. *Telematics and Informatics*, *32*(1), 67-78

Quantzig(2019),Social Media Analytics: Boosting Customer Engagement and Marketing Performance retrieved from https://www.businesswire.com/news/home/20190228005704/en/Social-Media-Analytics-Boosting-Customer-Engagement-Marketing

RafiqM(2017). The role of social media in Customer engagement retrieved from https://socialnomics.net/2017/07/14/the-role-of-social-media-in-customer-engagement/

Ray, G., Muhanna, W. A., & Barney, J. B. (2005). Information technology and the performance of the customer service process: A resource-based analysis. *MIS Quarterly*, 625-652.

Ruiz-Mafe, C., Martí-Parreño, J., & Sanz-Blas, S. (2014). Key drivers of consumer loyalty to Facebook fan pages. *Online Information Review*, *38*(3), 362-380.

SAP (2014), "Want Customers? Get Engaged", Retrieved fromhttp://www.prweb.com/releases/2014/08/prweb12115750.htm

SAS (2011), Improving Revenue and Customer Engagement With Social Media Analytics" , Retrieved from http://beta.content4demand.com/wp-content/uploads/2012/12/Improving-Renenue-And-Customer-Engagement-With-Social-Media-Analytics-WP-SAS.pdf

Setyani, V., Zhu, Y. Q., Hidayanto, A. N., Sandhyaduhita, P. I., & Hsiao, B. (2019). Exploring the psychological mechanisms from personalized advertisements to urge to buy impulsively on social media. *International Journal of Information Management*, *48*, 96-107.

Shiau, W., Dwivedi, Y. & Yang, H. (2017).Co-citation and cluster analyses of extant literature on social networks. *International Journal of Information Management*, 37(5), 390-399.

Shiau, W-L, Dwivedi, YK & Lai H-H (2018).Examining the core knowledge on Facebook. *International Journal of Information Management*, 43, 52-63.

Sigala, M., & Chalkiti, K. (2015).Knowledge management, social media and employee creativity. *International Journal of Hospitality Management*, *45*, 44-58.

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.

Sivarajah, U., Irani, Z., Gupta, S., & Mahroof, K. (2019). Role of big data and social media analytics for business to business sustainability: A participatory web context. *Industrial Marketing Management*.

Statista(2019), Most famous social network sites 2018, by active users, retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Clement,J. (2019), Number of monthly active Instagram users from January 2013 to June 2018 (in millions), retrieved from https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/

Stephenson A (2014), Engagement Part 2: How Engaged Are Your Customers? available at https://www.tlsasalestraining.com/content/engagement-part-2-how-engaged-are-your-customers

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

TCS(2014), “Customer Intelligence and Insights Solution for Retail, Digital Software & Solutions”, Retrieved from https://www.tcs.com/content/dam/tcs/pdf/product-details/DSS/customer-intelligence-retail.pdf

The Enterprise Strategy Group (2013) , “Social Media Marketing and Analytics, New Face of Customer Engagement”, Available at : http://www.oracle.com/us/corporate/analystreports/es-social-landscape-1936813.pdf

Thomson, M., MacInnis, D. J., & Park, C. W. (2005). The ties that bind: Measuring the strength of consumers’ emotional attachments to brands. *Journal of Consumer Psychology*, *15*(1), 77-91.

Umar, R. (2014),Social Media Analytics as a Business Intelligence Practice: Current Landscape & Future Prospects, *Journal of Internet Social Networking & Virtual Communities*, 2014, 1-12.

Venkatraman, N., & Ramanujam, V. (1987). Measurement of business economic performance: An examination of method convergence. *Journal of Management*, *13*(1), 109-122.

Wagner, C., &Majchrzak, A. (2006).Enabling customer-centricity using wikis and the wiki way. *Journal of Management Information Systems*, *23*(3), 17-43.

Yang, M., Kiang, M., Ku, Y., Chiu, C., & Li, Y. (2011). Social media analytics for radical opinion mining in hate group web forums. *Journal of Homeland Security and Emergency Management*, *8*(1).

Zaichkowsky, J. L. (1985). Measuring the involvement construct. *Journal of Consumer Research*, *12*(3), 341-352.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010).Social media analytics and intelligence. *IEEE Intelligent Systems*, *25*(6), 13-16.

APPENDIX
Measurement

Measures of the Social media analytics practices (SMAP)

Note: All items are measured using 5-point Likert-type scales with strongly disagree (1) to strongly agree (5)

Items marked by an asterisk (*) were removed from the final instrument.

Customer Management (CM)

SMAP/CM1 holistic and single view of the customer

SMAP/CM2 monitors customer activity and interactions

SMAP/CM3 track customer information in order to assess the lifetime value of each customer

SMAP/CM4 integrate customer information across customer contact points

Process Management (PM)

SMAP/PM1 Social media analytics strategy is connected to business objective/ outcomes

SMAP/PM2 Customer information is integrated across several functional areas

SMAP/PM3 proactively mitigate the risk

SMAP/PM4 analyse the competitors and their posting strategy, social media campaigns, and followers and learn their best practices

Performance Management (PEM)

SMAP/PEM1 Engagement or participation quantities

SMAP/PEM2 convert visitors into leads, and then into customers

SMAP/PEM3 Change in awareness or perceptions

SMAP/PEM4* return on investment (ROI)

SMAP/PEM5 Incremental revenue

SMAP/PEM6 Incremental sales

SMAP/PEM7* Prospects or leads generated

Measures of the Customer engagement (CE)

Note: All items are measured using 5-point Likert-type scales withNever (1) to all the time (5) *from the instrument.*

Involvement (IN) (Beatty & Talpade, 1994)

CE/IN1	average time spent on each page/ number of pages visited
CE/IN2	Visit Frequency
CE/IN3	visit services

Interaction (INT)

CE/INT1	Customers usually post /likes/share /comment/recommend/ blog about the products
CE/INT2	Facebook Wall Interaction
CE/INT3	performing the core user action

Intimacy (INM)

CE/INM1	customer service issues/requests are being handled
CE/INM2	Sentiments toward the subjects and the emotions expressed by the authors
CE/INM3	are engage in two-way dialogue and develop deeper relationships and a value add experience

Influence (INF)

CE/INF1*	Referring traffic
CE/INF2	customer Invite / Refer
CE/INF3	Customer Retweets

Measures of the Business Performance (BP)

Note: All items are measured using 5-point Likert-type scales with Decrease (1) to significant increase (5)

Items marked by an asterisk (*) were removed from the instrument.

Financial Performance (FP)	
BP/FP1*	Return on investment
BP/FP2	Profit margin on sales
Market Performance (MP)	
BP/MP1	Market Share
BP/MP2	Customer satisfaction
BP/MP3	Customer retention
BP/MP4	Sales growth

International Journal of Hospitality Management 72 (2018) 47–55

Contents lists available at ScienceDirect

International Journal of Hospitality Management

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

Digital marketing strategies, online reviews and hotel performance

Patrick De Pelsmacker[a,*], Sophie van Tilburg[b], Christian Holthof[b]

[a] University of Antwerp, Faculty of Applied Economics, Department of Marketing, Prinsstraat 13, 2000 Antwerpen, Belgium
[b] Artesis Plantijn University College, Faculty of Management and Communication, Department of Hotel Management, Meistraat 5, 2000 Antwerp, Belgium

ARTICLE INFO

Keywords:
Digital hotel marketing strategies
Volume and valence of online reviews
Hotel performance
Room occupancy
RevPar

ABSTRACT

We investigate to what extent digital marketing strategies (such as having a digital marketing plan, responsiveness to guest reviews, and monitoring and tracking online review information) influence hotel room occupancy and RevPar directly, and indirectly through the mediating effect of the volume and valence of online reviews they lead to, and to what extent this mechanism is different for different types of hotels in terms of star rating and independent versus chain hotels. The research was carried out in 132 Belgian hotels. The results indicate that review volume drives room occupancy and review valence impacts RevPar. Digital marketing strategies and tactics affect both the volume and valence of online reviews and, indirectly, hotel performance. This is more outspoken in chain hotels than in independent hotels, and in higher-star hotels than in lower-tier hotels.

1. Introduction

Electronic word-of-mouth (eWOM) is "all informal communications directed at consumers through Internet-based technology related to the usage or characteristics of particular goods and services, or their sellers" (Litvin et al., 2008). eWOM can take many forms, the most important one being online reviews. eWOM has a profound effect on attitudes and buying behavior of consumers and on commercial results in many product categories, such as books (Chevalier and Mayzlin, 2006), movies (Duan et al., 2008a; Liu, 2006), online games (Zhu and Zhang, 2010) and restaurants (Kim et al., 2016). eWOM appears to be particularly important for experience products. These are goods or services the quality of which cannot be judged easily prior to consumption, like hotels (Casalo et al., 2015). In such situations, the opinion of other consumers who post their experiences in online reviews, provides information from a source that is perceived as more independent and trustworth than company information (Zhao et al., 2015; Ye et al., 2011). In the travel industry, in the USA alone nearly two thirds of Web users relied on digital channels for travel information in 2013 (eMarketer, 2013). More than 74 percent of travelers use the comments of other consumers when planning trips (Gretzel and Yoo, 2008). Thus, online reviews are an important source of information for prospective hotel consumers, and they have an influence on trust and enjoyment (Sparks and Browning, 2011; Gretzel and Yoo, 2008), perceived credibility (Casalo et al., 2015; Mauri and Minazzi, 2013), hotel awareness (Vermeulen and Seegers, 2009), corporate reputation (Baka, 2016), attitudes (Casalo et al., 2015; Vermeulen and Seegers, 2009), hotel quality perceptions (Torres et al., 2015), booking intentions (Casalo et al., 2015; Ladhari and Michaud, 2015; Mauri and Minazzi, 2013; Sparks and Browning, 2011), hotel choice (Sparks and Browning, 2011; Vermeulen and Seegers, 2009), and willingness to pay (Nieto-García et al., 2017). As a result of this, online reviews also have an effect on hotel performance. Online reviews have been found to influence room occupancy, RevPar (revenue per available room), prices (Ögut and Tas, 2012; Ye et al., 2009, 2011) and market share (Duverger 2013).

Both the volume and the valence of online reviews affect consumer behavior (Kwok et al., 2017). Volume refers to the number of online reviews about a hotel in a given period; valence refers to the degree of positivity (rating) of these reviews (Blal and Sturman, 2014). More online comments have been found to lead to higher awareness (Zhao et al., 2015), and a better hotel performance (Viglia et al., 2016; Melián-González et al., 2013). The valence of online reviews also affects hotel performance. Ye et al. (2009, 2011) show that a 10% improvement in reviewers' rating can increase sales by 4.4%. Anderson (2012) reports that a 1-percent increase in a hotel's online reputation score leads up to a 0.89-percent increase in price, to a room occupancy increase of up to 0.54 percent, and to a 1.42-percent increase in RevPar. Viglia et al. (2016) report that a one-point increase in a hotel's review score is associated with an increase of 7.5 percentage points in the occupancy rate. Viglia et al. (2016) and Torres et al. (2015) find that both ratings and the number of reviews had a positive effect on online hotel bookings. Blal and Sturman (2014) demonstrate that, contrary to the number of reviews, there is a significant impact of ratings on

* Corresponding author.
E-mail addresses: Patrick.depelsmacker@uantwerpen.be (P. De Pelsmacker), sophie.vantilburg@ap.be (S. van Tilburg), christian.holthof@ap.be (C. Holthof).

https://doi.org/10.1016/j.ijhm.2018.01.003
Received 5 July 2017; Received in revised form 9 December 2017; Accepted 3 January 2018
0278-4319/

RevPar. However, very few studies have explored the potentially differential effects of review volume and valence on different indicators of hotel performance, such as room occupancy and RevPar.

An important question is what hotel marketing management can do to increase the volume and improve the valence of online reviews and, indirectly, hotel performance. Digital marketing strategies, such as closely monitoring and analyzing customer feedback (Torres et al., 2015), responding to customer feedback (Melian-Gonzalez and Bulchand-Gidumal, 2016; Sparks et al., 2016; Torres et al., 2015; Limb and Brymer, 2015; Wang et al., 2013; Levy et al., 2013; Chen and Xie, 2008), establishing a digital reputation management plan (Levy et al., 2013), monitoring and studying social media (Baka 2016; Levy et al., 2013) and integrating third-party review sites on the hotel website (Aluri et al., 2016) appear to drive online review volume and valence and/or hotel performance. However, Melian-Gonzalez and Bulchand-Gidumal (2016); Baka (2016) and Cohen and Olsen (2013) argue that further research is needed on how digital marketing strategies can enhance reviews and improve organizational performance.

Finally, what drives online reviews and how and to what extent these reviews impact hotel performance may be different for different types of hotels. Blal and Sturman (2014) and Phillips et al. (2017) argue that hotel characteristics are contextual factors that play an important moderating role in consumer behavior. Viglia et al. (2016) point out that belonging to a hotel chain or being higher-star-rated could be factors that increases hotel occupancy. However, only a few studies have focused on the moderating effect of hotel characteristics on the effect of online reviews on hotel performance, for instance unknown versus well-known hotels (Casalo et al., 2015), higher versus lower-tier hotels (star rating) (Blal and Sturman, 2014; Duverger, 2013), and chain versus independent hotels (Banerjee and Chua, 2016).

2. Purpose and contribution of the study

In the current study we try to partly fill three voids in the literature:

(1) How do volume and valence of online reviews affect different indicators of hotel performance, i.e. room occupancy and RevPar?
(2) Which digital marketing strategies drive hotel performance (room occupancy and RevPar) through the mediating role of the volume and valence of online reviews?
(3) Is this mechanism different for different types of hotels in terms of star rating and independent versus chain hotels?

Sainaghi (2010) proposes to measure hotel performances on the basis of three dimensions: financial (e.g. RevPar), operational (e.g. occupancy or repeat visit) and organizational (e.g. customer satisfaction). The current study uses room occupancy and RevPar as the dependent variables, representing an operational (quantity of bookings) and a financial (quality of bookings) dimension, respectively (Torres et al., 2015). An interesting question is to what extent digital marketing strategies and the volume and valence of reviews impact these two KPIs differentially (Blal and Sturman, 2014). In the current study, we answer the call for a more fine-grained analysis of the managerial and online review drivers of two different hotel performance indicators. The conceptual framework is shown in Fig. 1. Data were collected from 132 hotels in five tourist destinations in Flanders (Belgium), by means of a combination of a survey, a hotel website analysis, and online review data.

The study offers several insights into how hotel marketing works and provides guidelines for hotel marketing practice. Sainaghi (2010) distinguishes between external and internal determinants of hotel performance. The current study considers both. First, although the influence of online reviews (an external factor) on consumers' attitudes and behavior has been studied extensively, far less research has been reported on the influence of reviews on hotel performance. Studies that explore the effect of (digital) marketing strategies (an internal factor) on online reviews are also scarce (Sainaghi, 2010). Combining these two elements, the current study attempts to unravel the mechanism through which digital marketing strategies influence hotel performance, and the mediating role that volume and valence of online reviews play in this process. The current study also provides insights into the differential effects of digital marketing strategies and online reviews on hotel performance for different types of hotels, an important topic that only received scant attention (Sainaghi, 2010). The results of the current study can inform hotel marketing managers which elements of their digital marketing strategies to focus upon, what to expect from them in terms of their impact on different hotel performance indicators, and which online review elements should be monitored and taken into account in this process.

3. Literature review, research questions and hypotheses

3.1. The effect of online review volume and valence on hotel performance

The number of reviews a product/service receives from customers is one of the most critical review attributes (Duan et al., 2008b). Several studies have shown that more online reviews lead to a better business performance (Viglia et al., 2016; Kim et al., 2016; Zhu and Zhang, 2010; Duan et al., 2008b; Amblee and Bui, 2007; Chevalier and Mayzlin, 2006; Liu, 2006). Torres et al. (2015) and Ye et al. (2009) find that the number of reviews have a positive effect on online hotel bookings. Kim et al. (2015) report that the number of reviews has a significant effect on hotel revenues. Tuominen (2011) finds a positive relationship between the number of reviews and a hotel's RevPar and room occupancy. Viglia et al. (2016) report that, regardless the review score, the number of reviews has a positive effect with decreasing returns on the occupancy rate. The fact that review volume can positively affect business performance is attributed to the fact that reviews, positive or negative, are an indication of hotel popularity, increase consumers' awareness of the product, keep the product longer in people's consideration set, attract information seekers, reduce uncertainty and perceived risk, and trigger normative behavior ('go with the crowd') (Zhao et al., 2015; Viglia et al., 2014; Vermeulen and Seegers, 2009). This suggests that popularity *per se* has a strong relevance in terms of preferences (Viglia et al., 2016). Additionally, Torres et al. (2015) argue that, with greater number of reviews, the impact of extreme reviews is minimized.

Several studies have found that the valence of online reviews affects business performance. Positive consumer reviews increase business results, whereas negative online reviews decrease them (Anderson, 2012; Chevalier and Mayzlin, 2006). Positive comments can enhance the reputation of a company, while negative comments can reduce consumer interest in the company's products/services, which can affect its profits. Sparks and Browning (2011) argue that the overall valence of a set of hotel reviews affects customers' evaluations and trust and, consequently, booking intentions. Ye et al. (2009, 2011) show that positive online reviews can significantly increase the number of bookings in a hotel. They suggest that a 10% improvement in reviewers' rating can increase sales up to more than five percent. Limb and Brymer (2015) find that overall hotel ratings predict RevPar. Anderson (2012) reports that a 1% increase in a hotel's online reputation score leads to a room occupancy increase of up to 0.54 percent, and to a 1.42% increase in RevPar. Ögut and Tas (2012) show that a 1% increase in online customer rating increases sales per room up to 2.68% in Paris and up to 2.62% in London. In a study of 346 hotels in Rome, Viglia et al. (2016) found that a one-point increase in the review score is associated with a 7.5% point increase in the occupancy rate.

A few studies have assessed the impact of both the volume and the valence of online reviews on various indicators of hotel performance. In an online experiment, Nieto-Garcia et al. (2014) show a positive effect of review valence on willingness to pay for hotel accommodation, which is strengthened by online review volume. Viglia et al. (2014)

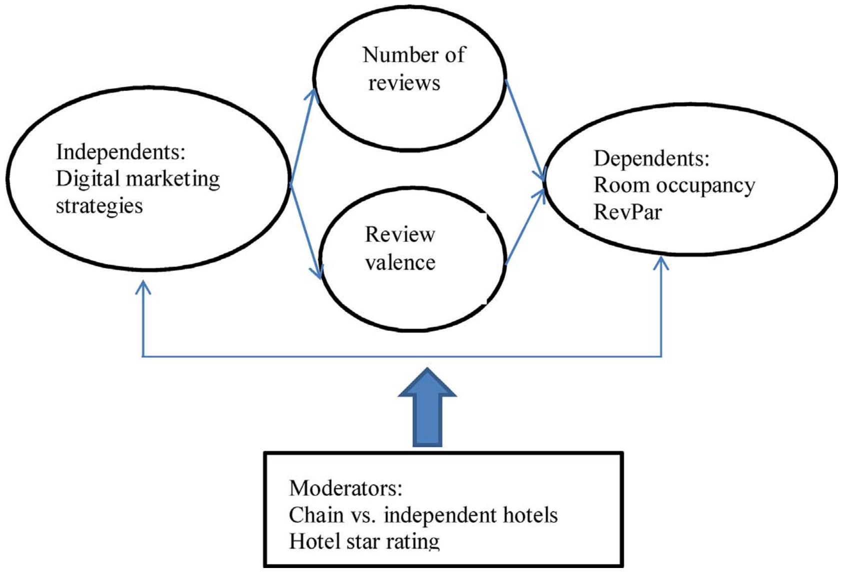


Fig. 1. Conceptual framework.

conducted an online conjoint experiment and found that consumers' preferences increased with both the number of reviews and the evaluation of the hotel. Torres et al. (2015) find that both ratings and the number of reviews on TripAdvisor had a positive effect on the average size of each online booking transaction. Each TripAdvisor star equated to an incremental $280 per booking transaction, and each review represented a total of $0.12 per booking transaction. Nieto-Garcia et al. (2014) find that both customer ratings and the number of reviews positively influenced profitability. Viglia et al. (2016) found a similar result for occupancy rate. On the other hand, Blal and Sturman (2014) and Limb and Brymer (2015) demonstrate that, contrary to the number of reviews, there is a significant impact of review ratings on RevPar. Using 56,284 hotel reviews posted for more than 1000 hotels listed on TripAdvisor, Xie et al. (2016) show that the effect of review valence lasts at least a couple of quarters, whereas that of review volume remains short-term. On the other hand, in the movie business, Duan et al. (2008a) found that the rating of online user reviews had no significant impact on movies' box office revenues, but were significantly influenced by the volume of online posting.

3.2. Digital marketing strategies, online reviews, and hotel performance

The effects of a hotel's digital marketing strategy on hotel performance, directly, or indirectly through its effect on online reviews, have only received scant attention in the academic literature (Cantallops and Salvi, 2014). Levy et al. (2013) and Melo et al. (2017) point out that hotels should establish a digital marketing plan, and that it is important for hotel managers to actively manage their online presence. In a digital hotel marketing plan, two main components can be distinguished. First, a hotel can actively use digital information in its marketing efforts in several ways, such as using information and metrics from review sites, providing a link to or integrating third-party reviews on its website, using track software to analyze reviews on OTA (Online Travel Agent) sites, or using OTAs management reports. Second, a hotel can have a conversation management strategy with its customers (for instance, responses to guest reviews, encouraging guests to post comments).

Several components of such a digital marketing plan have been explored in previous research. They are discussed hereafter. Information technologies (IT) have been recognized as one of the greatest forces causing change in the hotel industry (Law et al., 2013). Based on in-depth interviews with a group of 30 hotel managers, Melian-Gonzalez and Bulchand-Gidumal (2016) explore specific routes that IT can follow in order to improve hotel performance and argue that research is needed that clarifies how IT can improve this performance. Online feedback can help hotel managers track the attitudes, opinions, and satisfaction of guests and can serve as the basis for a series of management actions including responding to feedback, targeting investments in services that consumers would desire, and perpetuating positive actions. Hotel managers who place greater value on consumer-generated feedback are more likely to improve the perceived hotel quality (Torres et al., 2015).

Aluri et al. (2016) studied the influence of embedding social media channels on hotel websites on traveler behavior. They find that travelers exposed to a hotel website with embedded social media channels have higher levels of perceived informativeness, enjoyment, social interaction and satisfaction and, indirectly, purchase intention. Casalo et al. (2015) find that online ratings are considered more useful and credible when published by well-known online travel communities, such as TripAdvisor, leading to more favorable attitudes toward a hotel and higher booking intentions. Consequently, making these reviews explicitly and readily available on the hotel's website may have favorable effects on hotel performance. Melián-González et al. (2013) also argue that hoteliers should try to increase the number of reviews they receive and should therefore facilitate access to customer review sites.

The prominent role of social media necessitates that hotels also monitor online reviews for service recovery opportunities (Levy et al., 2013). Hotels are increasingly shifting from passive listening to active engagement through management responses. Online management responses are a form of customer relationship management (Gu and Ye, 2014). Management responses to a specific comment or a complaint in a consumer review show that hotel managers take their customers seriously, with the potential of improving customer reviews, customer satisfaction and, ultimately, hotel profitability (Sun and Kim, 2013; Chi and Gursoy, 2009). Various studies have explored the effect of responding to consumers' remarks, and especially negative remarks or complaints. Gu and Ye (2014) show that the satisfaction level of consumers who made complaints in their reviews increases after they received management responses. Xie et al. (2014) report a positive effect of the number of management responses to consumers' comments on hotel performance. They argue that these management responses will likely increase the consumer's likelihood of recommending the hotel, and will consequently influence the behavior of prospective customers.

Hotel management can respond to comments and complaints in different ways. Xie et al. (2017) report that providing timely responses enhances future financial performance, whereas providing responses by hotel executives and responses that simply repeat topics in the online review lowers future financial performance. A constructive response with a service recovery plan for negative reviews and a commitment to continuous effort for positive reviews drives purchase decisions by subsequent consumers. Functional staff/departments, rather than executives, should provide managerial responses because their operational insights allow them to better address consumer comments. Sparks et al. (2016) find that the provision of an online response, the timeliness of the response, and using a human voice rather than a professional one enhances trustworthiness and perceptions of caring. Levy et al. (2013) also suggest that the best response strategy is a positive, personalized response within a short period of time. On the basis of an experimental study with students, Min et al. (2015) conclude that using empathy in response to a negative review improved online ratings. The response was also rated more favorably when the response was more personal and less generic. Responses should thus include a strong signal that hotels do read the complaints, rather than repeatedly duplicating generic responses. On the other hand, and contrary to claims made in other studies, in the Min et al. (2015) study, the speed with which the hotel responded to a complaint did not influence the ratings. This may be explained by the fact that most people who read managerial responses are not complaining customers, but potential customers for whom the time element is less important.

All in all, previous studies have investigated the impact of digital strategies and customer reviews However, as Kwok et al. (2017) argue, much of this previous work mainly had a customer-centric perspective, focusing on customer decision making and customers responses such as trust and satisfaction, and there is an increasing research interest into examining the determinants of online reviews and the effect of online reviews on business performance. As Phillips et al. (2017) state, a question that previous research leaves open is which antecedent factors influence both room occupancy and RevPar, and how this is explained by the online reviews they generate. In the current study, we explore 10 aspects of a digital marketing strategy, and their effect on online reviews and, ultimately, hotel performance. We expect each of these digital strategies to have a positive impact:

H1. The following digital strategies have a positive effect on room occupancy and RevPar:

(a) Having a digital marketing plan,
(b) Using TripAdvisor information,
(c) Using TripAdvisor metrics,
(d) Using track software to analyze reviews on OTA sites
(e) Using OTAs management reports,
(f) Providing fast response to guest reviews,
(g) Providing personalized responses to guest reviews,
(h) Encouraging guests to post comments,.
(i) Providing a link to TripAdvisor,
(j) Integrating third-party reviews on its website.

These effect are mediated by both the volume and valence of online reviews.

3.3. Does this mechanism work differently for different types of hotels?

Several researchers argue that hotel characteristics are contextual factors that may play an important moderating role in consumer behavior, and call for further research into the effects of eWOM between different hotel categories (Phillips et al., 2017; Blal and Sturman, 2014; Cantallops and Salvi, 2014; Duverger, 2013).

Blal and Sturman (2014) report that review valence has a stronger effect on the RevPar of higher-tier hotels, while the volume of reviews has a greater effect on lower-tier hotels. The rating score effect on RevPar has little impact on the economy and midscale segments, while an increasing number of reviews actually has negative effects on higher-end hotels. These results apply equally to chain and independent hotels. They argue that, as room rates increase with the segment, the importance of the nature of the review on the purchasing decision increases. On the other hand, in lower-end segments, potential buyers need confirmation that the room is as advertised, and they rely more on the number of prior experiences. Similarly, Ögut and Tas (2012) find that the effect of customer ratings on sales was stronger for higher- star hotels and, in the same vein, Duverger (2013) concludes that lower-tiered hotels should not seek a high review rating, because it is mainly highly rated hotels that benefit from it.

Banerjee and Chua (2016) studied differences in online reviews for independent and chain hotels, and find review patterns to differ substantially between them. However, they did not explicitly study what drives these differences and how they relate to hotel performance. Compared with an unknown, unbranded independent hotel, a well-known hotel chain brand name may attenuate the influence of rating lists, because the consumer already has stable beliefs about it (Cantallops and Salvi, 2014). Indeed, Vermeulen and Seegers (2009) find that especially for lesser-known hotels reviews increase consumers' consideration of the hotel, and exposure to reviews has limited effect for well-known hotels.

In the current study, we explore the moderating role of hotel star rating and independent or chain hotels. Since previous research on the effect of hotel characteristics is scarce and contradictory, we propose the following research question:

RQ1. What is the moderating effect of hotel star rating and independent or chain hotels on the relationship between digital marketing strategies, volume and valence of online reviews, and hotel performance (room occupancy and RevPar)?

4. Method

4.1. Procedure and sample

The research was conducted in 2016 in the five officially recognized art cities in Flanders, Belgium: Antwerp, Bruges, Ghent, Mechelen and Leuven. On 31 December 2015, in those five cities, there were 224 licensed hotels. 37.5% were chain hotels, the other ones were independent. The Flemish government assigns a star rating to each hotel. Sixty-six hotels were 1–2-star rated, the others were 3–4-star rated, except for one that was 5 star rated. In January 2016, all hotels in these five cities received a paper survey in which, amongst others, the number of realized room nights, and digital marketing activities were measured. One hundred and thirty-two hotels returned a fully completed questionnaire, a response rate of almost 59%. In this sample, there were 23 1–2-star hotels and 109 3–4 star hotels. The 5-star hotel refused to cooperate for confidentiality reasons. Consequently, there are no five-star hotels in our sample. The sample contains 72 chain hotels and 60 independent hotels. Additionally, an analysis of the hotel websites was made in which elements of hotel online behavior were captured.

4.2. Measures

The dependent variables room occupancy (OCC) and RevPar were based on information reported in the survey. The list of independent variables (elements of a digital marketing strategy) was generated on the basis of in-depth interviews with 5 researchers from regional governmental or city tourism agencies, 2 representatives of hotel associations, 4 hotel tourism consultants, and 2 hotel managers (one 2-star and one 4-star hotel). The elements of digital marketing strategies are shown in Table 1. The first eight independent variables were measured in the survey; the last two were based on the hotel website analysis. The mediating variables, i.e. the number and valence of reviews in 2015 were made available by Olery, a company that tracks and analyses online reviews about hotels on more than 100 hotel review websites.

Table 1
Independent variables and frequencies per variable.

Variable	Variable name	Definition	Scale	0	1
Digital marketing plan	Digiplan	Whether or not the hotel has a digital marketing plan	0 (no) – 1 (yes)	88	26
Frequency TripAdvisor	Freq	The frequency with which hotel management uses TripAdvisor information	1 (a least weekly) – 0 (less frequently)	72	60
Metrics TripAdvisor	Metrics	Whether or not the hotel uses TripAdvisor metrics	0 (no) – 1 (yes)	40	63
Track software	Tracks	Whether or not the hotel uses track software to analyze reviews on OTA sites (e.g. Olery, Revinate, TrustYou)	0 (no) – 1 (yes)	85	44
Management reports	Reports	Whether or not the hotel uses OTA management reports	0 (no) – 1 (yes)	62	67
OTA 24 h	OTA24	Whether or not the hotel responds to guest remarks within 24 h	0 (no or hardly ever) – 1 (yes)	70	59
OTA personal answers	OTApers	Whether or not the hotel gives personalized answers to guest remarks (instead of standard answers)	0 (no or hardly ever) – 1 (yes)	43	86
OTA encourage posts	OTAencourage	Whether or not the hotel encourages guests to post comments	0 (no or hardly ever) – 1 (yes)	78	51
Link TripAdvisor	LinkTA	Whether or not the hotel has a link to TripAdvisor on its website	0 (no) – 1 (yes)	80	49
Integrated reviews	IntRev	Whether or not the hotel integrates commercial review sites' (e.g. TripAdvisor) reviews on its website	0 (no) – 1 (yes)	98	31

The valence of reviews is measured by means of the Guest Experience Index (GEI), Olery's proprietary confidential measure that is based on review ratings and sub-ratings (for attributes such as rooms, cleanliness, location and service), the integrity of the reviews (based on, amongst others, the credibility of the site and the frequency with which a person posts a review), review age, and a sentiment analysis of the reviews. GEI is expressed as a score between 0 (very bad) and 100 (outstanding). The moderators, i.e. the number of stars (1 or 2 vs. 3 or 4) and the hotel type (chain or independent) are based on official government data.

5. Analyses and results

The conceptual model in Fig. 1 was tested using Hayes' (2013) PROCESS macro for SPSS. Model 4 was used to test the basic mediation model. The Hayes procedure only allows models with one independent variable and one dependent variable. Therefore, in this first analysis, 20 models were tested, i.e. two (one per dependent variable) for each of the 10 independent variables. In each of these models, the number of reviews and the GEI were used as mediators. In Tables 2 and 3, the results of these estimations are shown. Only significant direct and indirect effects of the independent on one of the dependents are reported. Full statistical details can be obtained from the authors. In Table 2, the effects and their significance of each path between each of the model variables are reported. The independent variables are in the columns and the outcome variables in the rows. The direct effects of the digital

Table 3
Indirect effects of independents on dependents through mediation by number of reviews and GEI.

		Mediator = GEI	Mediator = Number
		Confidence interval	Confidence interval
Independent	Dependent		
Freq	OCC	[−0.0116; 0.0140]	[.0220; 0.0532]
Metrics	OCC	[−0.0207; 0.0010]	[.0078; 0.0425]
Digiplan	OCC	[−0.0087; 0.0064]	[.0093; 0.0720]
Tracks	OCC	[−0.0196; 0.0011]	[.0190; 0.0635]
Report	OCC	[−0.0129; 0.0020]	[.0114; 0.0475]
IntRev	OCC	[−0.0219; 0.001]	[.0029; 0.0495]
OTA24	OCC	[−0.0024; 0.0142]	[.0025; 0.0419]
Freq	RevPar	[2.8461; 12.6509]	[−5.2687; 2.4037]
IntRev	RevPar	[.1827;14.0967]	[−1.5606; 3.9503]

Notes: For variable names, please refer to Table 1. Cells of third and fourth rows are confidence intervals.

Table 2
Effects in the basic mediation model.

	Independent variables			*Independent variables*			*Independent variables*		
Outcome variables	Freq	GEI	Number	Metrics	GEI	Number	Digiplan	GEI	Number
GEI	4.466 (.001)			−2.644 (.041)			−0.113 (.938)		
Number	497.203 (< 0.001)			333.495 (.029)			491.003 (.003)		
OCC	0.054 **(.007)**	0.00002 (.855)	0.001 (< 0.001)	**−0.007 (.722)**	0.002 (.177)	0.001 (< 0.001)	**.017 (.465)**	0.002 (.240)	0.001 (< 0.001)
	Tracks	GEI	Number	Report	GEI	Number	IntRev	GEI	Number
GEI	−1.301 (.268)			−1.056 (.343)			−3.573 (.026)		
Number	536.973 (< 0.001)			326.407 (.010)			378.222 (.010)		
OCC	**.060 (.003)**	0.003 (.030)	0.001 (< 0.001)	**−0.004 (.838)**	0.003 (.072)	0.001 (< 0.001)	**.060 (.006)**	0.002 (.076)	0.001 (< 0.001)
	OTA24	GEI	Number	Freq	GEI	Number	IntRev	GEI	Number
GEI	0.775 (.489)			4.466 (.001)			−3.573 (.026)		
Number	279.833 (.027)			497.203 (< 0.001)			378.222 (.010)		
OCC	**−0.005 (.783)**	0.003 (.066)	0.001 (< 0.001)						
RevPar				**17.960** (< **.001**)	1.685 (< 0.001)	−0.002 (.607)	**3.914 (.481)**	1.995 (< 0.001)	0.002 (.580)

Notes: For variable names, please refer to Table 1. Cells are path coefficients (significance).

strategies on either OCC or RevPar are in bold. Table 3 reports the indirect effects of the independents on the dependents, through the mediation role of both the number of reviews and GEI. Each row refers to one model estimation. In the third and fourth columns of this table, confidence intervals are given. When a confidence interval does not contain zero, the indicated indirect effect is statistically significant ($p < 0.05$).

The frequency of TripAdvisor information used, using track software and integrating commercial review sites' reviews on the hotel website have both a direct and an indirect effect on room occupancy, and thus their positive effect is partly mediated by the number of reviews these digital strategies generate. Using TripAdvisor metrics, having a digital marketing plan, using management reports and answering guest comments within 24 h only have an indirect effect on room occupancy, and thus the effect of these digital strategies on room occupancy is fully mediated by the number of reviews these strategies generate. H1a,b,c,d,e,f,j are supported as far as room occupancy and the mediating role of review volume are concerned. None of these effects are partly or fully mediated by GEI. H1 is thus not supported with respect to the mediating role of GEI on occupancy. A personalized response to guest remarks, encouraging OTA reviews and a link to TripAdvisor on the hotel website have neither a direct nor an indirect effect (through online reviews) on room occupancy. H1 g,h,i are not supported for room occupancy. The frequency of using TripAdvisor information has a direct positive effect on RevPar, and an indirect positive effect through GEI. Integrating reviews on the hotel website also has a positive indirect effect on RevPar, through its beneficial effect on GEI. H1b,j are supported as far as RevPar and the mediating role of review valence are concerned. The number of reviews does not mediate these effects on RevPar. H1 is thus not supported with respect to the mediating role of review volume on RevPar. H1a,c,d,e,f,g,h are not supported for RevPar.

In the second set of analyses, we answer RQ1 by testing the moderating effect of the number of stars (1 or 2 vs. 3 or 4) and the hotel type (chain or independent) on the mediation process documented in the previous analysis, using Hayes' PROCESS macro 59. We only made these moderation analyses on mediation models that showed significant effects (the 9 models reported in Table 3). We only report significant moderation effects. Full statistical details can be obtained from the authors. The main indicator for judging the meaningfulness of a moderation effect is the difference in conditional effect sizes (as detailed in Table 4 for RevPar and Table 5 for room occupancy) for the two different values of the moderators. Additionally, a clear indication of moderation would be that there is a significant effect for one of the values of the moderator, but not for the other. If a confidence interval in Tables 4 and 5 contains zero, the conditional effect is not significant for that value of the moderator. We have used these criteria to arrive at our conclusions. In Tables 4 and 5, the first column indicates the number of the analysis. The second column shows the two levels of the moderator. The next three columns show the direct effects of the independent on the dependent, for the two values of the moderator. The last two columns of each table show the effect sizes and the confidence intervals of the indirect effects through the mediator.

Table 4 shows that there is a direct positive effect of frequency on RevPar for independent hotels, but not for chain hotels. However, the indirect effect through generating a higher GEI score is stronger for chain hotels than for independent hotels (analysis 1). Both the direct and indirect (through GEI) effects of frequency on RevPar are stronger for higher rated hotels (analysis 2). The indirect effect (through GEI) of integrating reviews on the hotel website is negative for chain hotels, and insignificant for independent hotels (analysis 3). The indirect effect of Intrev on RevPar is negative for low star hotels, and not significant for high star ones (analysis 4).

Table 5 indicates that the direct effect of frequency on room occupancy is positive for independent hotels and not significant for chain hotels, but the positive indirect effect through the number of posted reviews is only significant for chain hotels and not for independent hotels (analysis 5). There is only a direct effect of frequency on room occupancy for high star hotels, but its indirect effect through the number of online reviews is stronger for low star hotels (analysis 6). There is no direct effect of the use of metrics on room occupancy, and the indirect effect is only significant for chain hotels (analysis 7) and high star hotels (analysis 8). The indirect effect (through the number of online reviews) of the use of track software on room occupancy is stronger for independent hotels (analysis 9), and both the direct and indirect effects are only significant for high-star hotels (analysis 10). The indirect effect of responding to guest reviews within 24 h (through the number of reviews) on room occupancy is only significant for high-star hotels. There is no direct effect of this strategy on room occupancy (analysis 11). The indirect effect of using reports on room occupancy (through the number of reviews) is only significant for chain hotels (analysis 12). The direct effect of integrating third party reviews on the hotel website is only significant for chain hotels (analysis 13). The indirect effect of having a digital marketing plan (through the number of reviews) on room occupancy is only significant for high-star hotels (analysis 14).

6. Conclusions and discussion

Digital marketing strategies such as having a digital marketing plan, the frequency of TripAdvisor information used, using TripAdvisor metrics, using management reports, answering guest reviews within 24 h, using track software, and integrating commercial review sites' reviews on the hotel website, all appear to affect room occupancy favorably, partly or fully because they lead to more reviews, and not because they increase review valence (GEI). Stated differently, the positive effect of the use of these strategies on room occupancy is partly or fully mediated by the number of reviews these activities generate. These results confirm previous findings on the role of review volume on room occupancy (Torres et al., 2015; Tuominen, 2011; Ye et al., 2009). Review valence does not affect room occupancy. This contradicts earlier findings by Ye et al. (2009) and Anderson (2012).

Table 4
Conditional direct and indirect effects of the moderators Hotel type and star rating; independents: Freq and Intrev; mediator: GEI, Dependent: RevPar.

Analysis	Hotel type	Direct effects of Freq on RevPar			Indirect effect Mediator = GEI	
		Effect	*Sign.level*	*Conf. interval*	*Effect*	*Conf. interval*
1	Chain	10.064	0.058	[−0.3245;20.4533]	12.944	[4.1721;23.5273]
	Independent	15.863	0.019	[2.6130;29.1128]	6.709	[2.3150;14.1122]
2	Low star	11.820	0.048	[.0846;23.5551]	2.041	[−3.7973;6.4393]
	High star	14.930	< 0.001	[6.7945;23.0645$	9.7548	[3.8583;18.1402]
		Direct effects of Intrev on RevPar				
3	Chain	−2.882	0.637	[−14.9286;9.1646]	−11.020	[−20.1778;−2.4373]
	Independent	2.711	0.922	[−52.2884;57.7101]	6.182	[−5.9572;22.4701]
4	Low star	−2.206	0.693	[−13.2527;8.8404]	−6.028	[−19.2386;−0.0475]
	High star	5.053	0.333	[−5.2373;15.3435]	−4.684	[−13.6720;2.3914]

Table 5
Conditional direct and indirect effects of the moderators Hotel type and star rating; independents: Freq, Metrics, Tracks, OTA24, Report, Intrev and Digiplan; mediator: Number, Dependent: OCC.

Analysis	Hotel type	Effect	Sign.level	Conf. interval	Indirect effect Mediator = Number	
		Direct effects of Freq on OCC			*Effect*	*Conf. interval*
5	Chain	0.024	0.321	[−0.0239;.0724]	0.033	[.0196;.0555]
	Independent	0.068	0.026	[.0082;.1281]	0.015	[−0.0177;.0606]
6	Low star	−0.020	0.772	[−0.1586;.1180]	0.073	[.0239;.1707]
	High star	0.071	0.002	[.0271;.1147]	0.035	[.0182;.0548]
		Direct effects of Metrics on OCC				
7	Chain	−0.015	0.491	[−0.0598;.0289]	0.030	[.0132;.0507]
	Independent	0.005	0.889	[−0.0655;.0755]	−0.008	[−0.0567;.0340]
8	Low star	0.025	0.685	[−0.0975;.1478]	0.017	[−0.0121;.1131]
	High star	−0.014	0.469	[−0.0530;.0246]	0.026	[.0049;.0496]
		Direct effects of Tracks on OCC				
9	Chain	0.036	0.090	[−0.0056;.0767]	0.022	[−0.0013;.0503]
	Independent	0.067	0.136	[−0.0215;.1559]	0.046	[.0016;.1308]
10	Low star	0.069	0.228	[−0.0437;.1820]	0.026	[−0.0108;.1331]
	High star	0.049	0.008	[.0131;.0854]	0.046	[.0262;.0735]
		Direct effects of OTA24 on OCC				
11	Low star	−0.081	0.185	[−0.2007;.0392]	0.011	[−0.0864;.1626]
	High star	0.006	0.767	[−0.0357;.0483]	0.033	[.0101;.0579]
		Direct effects of Report on OCC				
12	Chain	−0.002	0.916	[−0.0399;.0359]	0.027	[.0081;.0460]
	Independent	−0.014	0.658	[−0.0785;.0498]	0.008	[−0.0253;.0427]
		Direct effects of IntRev on OCC				
13	Chain	0.047	0.004	[.0153;.0784]	0.013	[−0.0094;.0332]
	Independent	0.015	0.848	[−0.1396;.1697]	−0.012	[−0.0943;.0667]
		Direct effects of Digiplan on OCC				
14	Low star	0.020	0.841	[−0.1802;.2208]	0.017	[−0.0174;.1220]
	High star	0.010	0.658	[−0.0336;.0531]	0.048	[.0171;.0957]

RevPar is affected by digital strategies to a more limited extent, i.e. only by the frequency of using TripAdvisor information and by integrating reviews on the hotel website. Both effects are mediated by review valence. This finding is in line with previous studies (e.g., Limb and Brymer, 2015; Anderson, 2012; Ögut and Tas, 2012). The fact that not the number of reviews, but their valence affects RevPar, is a confirmation of the findings of Blal and Sturman (2014) and Limb and Brymer (2015), but contradicts the findings of Torres et al. (2015) and Nieto-Garcia et al. (2014) in that the latter find an effect of both review volume and valence on hotel profitability.

All in all, most components of a digital marketing strategy considered in the current study affect hotel performance, partly or fully through the effect they have on the volume and/or valence of online reviews. However, this is more the case for room occupancy than for RevPar, and review volume mediates the effect of digital strategies on room occupancy, while review valence does so for the effect of strategies on RevPar. These results confirm Blal and Sturman's (2014) claim that volume and valence of online reviews influence hotel performance parameters differently. The results also confirm the crucial role of IT strategies and, to a lesser extent, the importance of responsiveness and service recovery. As to the former, the present findings support the need for a formal digital marketing plan (Levy et al., 2013) and the impact of embedding social media channels and integrating reviews on the hotel website (Aluri et al., 2016; Melo et al., 2017). As to the need for responding to guests' comments, only the speed of the response to comments seems to matter. This confirms previous findings (Xie et al., 2017; Sparks et al., 2016; Levy et al., 2013), although our results are not consistent with Min et al.'s (2015) conclusion.

Remarkably, a link to TripAdvisor on the hotel website, personalized response to guest remarks, and encouraging guests to post reviews, have no effect on either room occupancy or RevPar. The fact that a link to TripAdvisor does not stimulate reviews and improve hotel performance, contradicts the findings of Casalo et al. (2015) that well-known online communities lead to better attitudes and booking intentions. The fact that a personalized response and encouraging reviews from guests do not have an effect on reviews and hotel performance, contradicts several previous studies (Levy et al., 2013; Min et al., 2015; Sparks et al., 2016; Xie et al., 2017). This is an unexpected result that requires further study.

Of all digital strategies considered in the current study, the frequency of using TripAdvisor information and integrating third party reviews on the hotel website seems to be most important, since they have an impact on both the number and the valence of online reviews, which in turn leads to both a higher room occupancy and RevPar.

These mediation processes are moderated by the type (chain or independent) and the star rating of the hotel. A rather consistent result is that the effect of a number of digital strategies appears to be stronger for higher-star hotels, either in their direct effect on room occupancy or indirectly through their effect on the number of reviews posted, or both. This is the case for having a digital marketing plan, using metrics and track software, and responding to guest reviews within 24 h. Furthermore, both the direct and indirect (through GEI) effect of the frequency of use of TripAdvisor information on RevPar is stronger for higher-rated hotels, and the indirect effect (through GEI) of integrating third-party review sites is negative for lower-star hotels, and not significant for higher-star ones. These results confirm the findings of Ögut and Tas (2012) and Duverger (2013), and partly those of Blal and Sturman's (2014). The latter find that review valence had a stronger effect on the RevPar of high-star hotels than on economy and midscale segments. However, their finding that review volume drives RevPar of lower-end hotels is not confirmed, quite the contrary, since our findings suggest that also the effect of review volume on room occupancy plays a stronger role for higher rated hotels.

The indirect effect (trough the number of reviews) of online strategies on room occupancy is generally stronger for chain hotels than for independent hotels. This is the case for the frequency of using TripAdvisor information, using metrics and reports, and integrating third party reviews on the hotel website. The indirect effect of the frequency of using TripAdvisor information on RevPar through generating a higher GEI score, is stronger for chain hotels than for independent hotels as well. However, the effect of the use of track software on room occupancy rate is stronger for independent hotels, and so is the direct effect of frequently using TripAdvisor information on RevPar. Responding quickly to guest reviews has a negative effect on room occupancy for chain hotels, but not for independent hotels, and the effect of integrating third-party reviews on the hotel website has a negative effect on chain hotels, and no effect on independent ones. Some of these results contradict Cantallops and Salvi's (2014) and Vermeulen and Seegers's (2009) findings, which are explained by the assumption that chain hotels have well-known brand names and are more familiar to the traveler, and that this may diminish the influence of online reviews. We believe that our results could be explained by the fact that chain hotels may have a more professional and sophisticated, and thus more powerful, digital marketing strategy, which leads to a greater impact of digital tactics on online reviews and hotel performance.

7. Managerial implications

The managerial implications of our study are that hotel management should devote considerable attention to both the number and the valence of reviews about their hotel, and should develop an extensive digital marketing strategy that has a profound impact on these reviews and, directly or indirectly, on hotel performance. The first step in such a digital marketing strategy is to have a digital marketing plan that provides for online hotel presence, tracking and monitoring online reviews, and quick response to customer comments. Indeed, many components of such a plan have a significant impact on the volume and/or the valence of online reviews, and on hotel performance in terms of room occupancy or RevPar. This is especially true for chain hotels and 3–4 star hotels, for which the impact of digital strategies and tactics is, generally speaking, more outspoken than for independent or lower-tier hotels. Especially the frequency of using TripAdvisor information (reports and metrics) and integrating third party reviews on the hotel website is crucial, since these tactics increase both the volume of and the appreciation in online reviews and, as such, indirectly influence both room occupancy and RevPar positively. The TripAdvisor Management Dashboard is an analytics service that summarizes a hotel's performance on TripAdvisor. Hotels can use the data and information to track how they are engaging with customers and guests online, target areas for improvement and make informed, decisions. The dashboard provides, amongst others, reports on a hotel's total reviews and popularity ranking over time and relative to the hotel's competitors in the same geographical region, latest review activity and top comments from reviews, number of traveler and hotel-submitted photos, and the number of visitors viewing photos, most viewed competitors, the countries generating the most traffic to the hotel's TripAdvisor page, trends over time, and performance metrics. TripAdvisor reports provide, amongst others, business trends, risk factors, financial data, and results of operations (www.tripadvisor.com). Additionally, hotels that strive for a higher room occupancy should aim at increasing the volume of online reviews This online review volume can be increased by increasing the frequency of using TripAdvisor metrics, by using track software and management reports, and by answering guest comments within 24 h. Hotels that want to focus upon RevPar need to improve the valence of online reviews.

8. Limitations and future research

Our study has some limitations that offer opportunities for further research. First, the current study was carried out in 132 hotels in five Belgian cities. Our findings should be corroborated in different countries and contexts and in larger samples. There were no luxury (5-star) hotels in our sample. Further research could also focus on this special hotel category. Second, two contextual variables are taken into account (independent versus chain and star rating), but different contextual factors could be considered, such as, for instance, the size of the hotel, the region in which the hotel is situated, and the type of visitors (e.g. business, leisure). In any case, the scarcity of studies on the influence of contextual factors on the effect of online strategies on reviews and hotel performance necessitates further research in this area. Third, volume and valence are the most frequently studied aspects of online reviews, but also other elements could be taken into account, such as the variance of reviews, the percentage of negative reviews, the topic of the reviews (hotel and service attributes, for instance), the degree of negativity and positivity of reviews, reviewer characteristics (demographics, reputation expertise, experience), etc. (Kwok et al., 2017). Next, a number of digital strategies and tactics have been taken into account, but there could be other factors that stimulate the generation of reviews, such as service characteristics, hotel amenities, staff behavior, location etc. Their relative importance and how management strategies can enhance or attenuate their effects should be studied. Finally, the effect of managerial responses to customer comments should be studied further. Most studies to date investigate the effect of managerial responses on customer trust, concern, satisfaction, and attitudes. However, this research should be taken one step further, and explore the effect of managerial responses on hotel performance, and the mediating role of customer reactions in this process.

References

Ögut, H., Tas, B.K.O., 2012. The influence of internet customer reviews on the online sales and prices in hotel industry. Serv. Ind. J. 32 (2), 197–214.
Aluri, A., Slevitch, L., Larzelere, R., 2016. The influence of embedded social media channels on travelers' gratifications, satisfaction, and purchase intentions. Cornell Hosp. Q. 57 (3), 250–267.
Amblee, N., Bui, T.X., 2007. The impact of electronic-word-of-mouth on digital microproducts: an empirical investigation of Amazon shorts. In: Österle, H., Acton, T., Conboy, K. (Eds.), Proceedings of the 15th European Conference on Information Systems. AIS Electronic Library, St. Gallen, pp. 36–47.
Anderson, C., 2012. The impact of social media on lodging performance. Cornell Hosp. Rep. 12 (15), 6–11.
Baka, K., 2016. The becoming of user-generated reviews: looking at the past to understand the future of managing reputation in the travel sector. Tour. Manag. 53, 148–162.
Banerjee, S., Chua, A.Y.K., 2016. In search of patterns among travelers' hotel ratings in TripAdvisor. Tour. Manag. 53, 125–131.
Blal, I., Sturman, M.C., 2014. The differential effects of the quality and quantity of online reviews on hotel room sales. Cornell Hosp. Q. 55 (4), 365–375.
Cantallops, A.S., Salvi, F., 2014. New consumer behavior: a review of research on eWOM and hotels. Int. J. Hosp. Manag. 36, 41–51.
Casalo, L.V., Flavian, C., Guinaliu, M., Ekinci, Y., 2015. Do online hotel rating schemes influence booking behaviors? Int. J. Hosp. Manag. 49, 28–36.
Chen, Y., Xie, J., 2008. Online consumer review: word-of-mouth as a new element of marketing communication mix. Manag. Sci. 54 (3), 477–491.
Chevalier, J.A., Mayzlin, D., 2006. The effects of word of mouth on sales: online book reviews. J. Mark. Res. 43 (3), 345–354.
Chi, C.G., Gursoy, D., 2009. Employee satisfaction, customer satisfaction, and financial performance: an empirical examination. Int. J. Hosp. Manag. 28 (2), 245–253.
Cohen, J.F., Olsen, K., 2013. The impacts of complementary information technology resources on the service-profit chain and competitive performance of South African hospitality firms. Int. J. Hosp. Manag. 34, 245–254.
Duan, W., Gu, B., Whinston, A.B., 2008a. Do online reviews matter?—An empirical investigation of panel data. Decis. Support Syst. 45 (4), 1007–1016.
Duan, W., Gu, B., Whinston, A.B., 2008b. The dynamics of online word of mouth and product sales: an empirical investigation of the movie industry. J. Retail. 84 (2), 233–242.
Duverger, P., 2013. Curvilinear effects of user-generated content on hotels' market share: a dynamic panel-data analysis. J. Travel Res. 52 (4), 465–478.
Gu, B., Ye, Q., 2014. First step in social media: measuring the influence of online management responses on customer satisfaction. Prod. Oper. Manag. 23, 570–582.
Hayes, A.F., 2013. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach. Guilford Press, New York/London.
Kim, W.G., Lim, H., Brymer, R.A., 2015. The effectiveness of managing social media on hotel performance. Int. J. Hosp. Manag. 44, 165–171.
Kim, W.G., Lim, J., Brymer, R.A., 2016. The impact of social media reviews on restaurant performance: the moderating role of excellence certificate. Int. J. Hosp. Manag. 55, 41–51.
Kwok, L., Xie, K.L., Richards, T., 2017. Thematic framework of online review research: a systematic analysis of contemporary literature on seven major hospitality and tourism journals. Int. J. Contemp. Hosp. Manag. 29 (1), 307–354.
Ladhari, R., Michaud, M., 2015. eWOM effects on hotel booking intentions,attitudes, trust: and website perceptions. Int. J. Hosp. Manag. 46, 36–45.
Law, R., Leung, D., Au, N., Lee, A., 2013. Progress and development of information technology in the hospitality industry evidence from Cornell Hospitality Quarterly. Cornell Hosp. Q. 54 (1), 10–24.
Levy, S.E., Duan, W., Boo, S., 2013. An analysis of one-star online reviews and responses

in the Washington, D.C. lodging market. Cornell Hosp. Q. 54 (1), 49–63.
Litvin, S.W., Goldsmith, R.E., Pan, B., 2008. Electronic word-of-mouth in hospitality and tourism management. Tour. Manag. 29, 458–468.
Liu, Y., 2006. Word of mouth for movies: its dynamics and impact on box office revenue. J. Market. 70 (July), 74–89.
Mauri, A., Minazzi, R., 2013. Web reviews influence on expectations and purchasing intentions of hotel potential customers. Int. Jo. Hosp. Manag. 34, 99–107.
Melián-González, S., Bulchand-Gidumal, J., González López-Valcárcel, B., 2013. Online customer reviews of hotels: as participation increases, better evaluation is obtained. Cornell Hosp. Q. 54 (3), 274–283.
Melian-Gonzalez, S., Bulchand-Gidumal, J., 2016. A model that connects information technology and hotel performance. Tour. Manag. 53, 30–37.
Melo, A.J., Hernández-Maestro, R.M., Muñoz-Gallego, P.A., 2017. Service quality perceptions, online visibility, and business performance in rural lodging establishments. J. Travel Res. 56 (2), 250–262.
Min, H., Lim, Y., Magnini, V.P., 2015. Factors affecting customer satisfaction in responses to negative online hotel reviews: the impact of empathy, paraphrasing, and speed. Cornell Hosp. Q. 56 (2), 223–231.
Nieto-García, M., Munoz-Gallego, P.A., González-Benito, O., 2017. Tourists' willingness to pay for an accommodation: the effect of eWOM and internal reference price. Int. J. Hosp. Manag. 62, 67–77.
Nieto-Garcia, M., Hernandez-Maestro, R.M., Munoz-Gallego, P.A., 2014. Marketing decisions, customer reviews, and business performance: the use of the Top rural website by Spanish rural lodging establishments. Tour. Manag. 45, 115–123.
Phillips, P., Barnes, S., Zigan, K., Schegg, R., 2017. Understanding the impact of online reviews on hotel performance: an empirical analysis. J. Travel Res. 56 (2), 235–249.
Sainaghi, R., 2010. Hotel performance: state of the art. Int. J. Contemp. Hosp. Manag. 22 (7), 920–952.
Sparks, B.A., Browning, V., 2011. The impact of online reviews on hotel booking intentions and perception of trust. Tour. Manag. 32, 1310–1323.
Sparks, B.A., So, K., Bradley, G.L., 2016. Responding to negative online reviews: the effects of hotel responses on customer inferences of trust and concern. Tour. Manag. 53, 74–85.
Sun, K.-A., Kim, D.-Y., 2013. Does customer satisfaction increase firm performance?An application of American Customer Satisfaction Index (ACSI). Int. J. Hosp. Manag. 35, 68–77.
Torres, E.N., Singh, D., Robertson-Ring, A., 2015. Consumer reviews and the creation of booking transaction value: lessons from the hotel industry. Int. J. Hosp. Manag. 50, 77–83.
Tuominen, P., 2011. The influence of TripAdvisor consumer-generated travel reviews on hotel performance. University of Hertfordshire Business School Working Paper, Presented at the 19th Annual Frontiers in Service Conference 1–11.
Vermeulen, I.E., Seegers, D., 2009. Tried and tested: the impact of online hotel reviews on consumer consideration. Tour. Manag. 30, 123–127.
Viglia, G., Furlan, R., Ladron-de-Guevara, A., 2014. Please, talk about it! When hotel popularity boosts preferences. Int. J. Hosp. Manag. 42, 155–164.
Viglia, G., Minazzi, R., Buhalis, D., 2016. The influence of e-word-of-mouth on hotel occupancy rate. Int. J. Contemp. Hosp. Manag. 28 (9), 2035–2051.
Wang, W., Miao, L., Huang, Z.J., 2013. Customer engagement behaviors and hotel responses. Int. J. Hosp. Manag. 33, 316–330. www.tripadvisor.com.
Xie, L., Zhang, Z., Zhang, Z., 2014. The business value of online consumer reviews and management response to hotel performance. Int. J. Hosp. Manag. 43, 1–12.
Xie, K.L., Chen, C., Wu, S., 2016. Online consumer review factors affecting offline hotel popularity: evidence from dvisor. J. Travel Tour. Mark. 33 (2). http://dx.doi.org/10.1080/10548408.2015.1050538.
Xie, K.L., So, K., Wang, W., 2017. Joint effects of management responses and online reviews on hotel financial performance: a data-analytics approach. Int. J. Hosp. Manag. 62, 101–110.
Ye, Q., Law, R., Gu, B., 2009. The impact of online user reviews on hotel room sales International. J. Hosp. Manag. 28, 180–182.
Ye, Q., Law, R., Gu, B., Chen, W., 2011. The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Comput. Hum. Behav. 27, 634–639.
Zhao, X., Wang, L., Guo, X., Law, R., 2015. The influence of online reviews to online hotel booking intentions. Int. J. Contemp. Hosp. Manag. 27 (6), 1343–1364.
Zhu, F., Zhang, X., 2010. Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics. J. Mark. 74, 133–148.
eMarketer, 2013. Metasearch Growth Reflects Travelers' Appetite for Information. available at: www.emarketer.com/Article/Metasearch-Growth-Reflects-Travelers-Appetite-Information/1009853 (Accessed 18 June 2013).

Journal of Innovation & Knowledge 6 (2021) 92–102

Journal of Innovation & Knowledge

https://www.journals.elsevier.com/journal-of-innovation-and-knowledge

Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics

Jose Ramon Saura

Department of Business Economics, Rey Juan Carlos University, Madrid, Spain

A R T I C L E I N F O

Article history:
Received 30 March 2020
Accepted 3 August 2020
Available online 15 August 2020

JEL classification:
M15
M31

Keywords:
Data Sciences
Digital Marketing
Knowledge discovery
Literature review
Data Mining

A B S T R A C T

In the last decade, the use of Data Sciences, which facilitate decision-making and extraction of actionable insights and knowledge from large datasets in the digital marketing environment, has remarkably increased. However, despite these advances, relevant evidence on the measures to improve the management of Data Sciences in digital marketing remains scarce. To bridge this gap in the literature, the present study aims to review (i) methods of analysis, (ii) uses, and (iii) performance metrics based on Data Sciences as used in digital marketing techniques and strategies. To this end, a comprehensive literature review of major scientific contributions made so far in this research area is undertaken. The results present a holistic overview of the main applications of Data Sciences to digital marketing and generate insights related to the creation of innovative Data Mining and knowledge discovery techniques. Important theoretical implications are discussed, and a list of topics is offered for further research in this field. The review concludes with formulating recommendations on the development of digital marketing strategies for businesses, marketers, and non-technical researchers and with an outline of directions of further research on innovative Data Mining and knowledge discovery applications.

Introduction

Since the beginning of the 21st century, both Digital Marketing (DM) and Data Sciences (DS) have remarkably evolved in terms of use and profitability (Tiago & Veríssimo, 2014). This has led to the emergence of a digital ecosystem, which connects users 24/7 and which has shaped users' new habits and behaviors (Mayer-Schönberger & Cukier, 2013).

DM is defined as a set of techniques developed on the Internet with to persuade users to buy a product or service (Avery, Steenburgh, Deighton, & Caravella, 2012). Today, the daily roadmap for companies that operate on the Internet includes techniques such as Search Engine Optimization (SEO), i.e. optimization of search results from major search engines; Search Engine Marketing (SEM) or programmatic advertising, i.e. strategies to sponsor ads in search engines or in advertising space on banners in websites; as well as Social Media Marketing (SMM), i.e. strategies of interacting with users on social networks through social ads (Lies, 2019; Palos-Sanchez, Saura, & Martin-Velicia, 2019).

In recent years, DM has spurred a considerable research interest among scholars (Kannan, 2017). For instance, Rogers and Sexton (2012) sought to understand the key ways to improve profitability or ROI (Return of Investment) in DM. Furthermore, Kumar et al. (2013) measured the influence of data on the DM ecosystem. Likewise, Saura, Palos-Sánchez, and Cerdá Suárez (2017) identified the metrics to measure the efficiency of each of the DM actions developed by a company on the Internet.

Numerous studies demonstrated that a key way to increase the effectiveness of DM strategies is the application of DS techniques in this industry (Braverman, 2015; Dremel, Herterich, Wulf, & Vom Brocke, 2020; Sundsøy, Bjelland, Iqbal, & de Montjoye, 2014). For example, Kelleher and Tierney (2018) argued that DS can increase the effectiveness of DM by improving issues such as (i) companies' management of the information collected from users; (ii) the type and source of data from the companies' datasets, and (iii) application of new data analysis and innovative techniques to create knowledge (Palacios-Marqués et al., 2016).

Furthermore, Fan, Lau, and Zhao (2015) and Saura and Bennett (2019) underscored the importance of several important aspects, such as the type of data collected from different online sources, purchases made by users, and their digital habits or behaviors. Likewise, Wedel and Kannan (2016) demonstrated that, in order to increase the chances of success on digital and social media platforms, companies should identify unsuspected patterns using Artificial Intelligence (AI) or Machine Learning (ML) techniques. Accordingly, the DM industry has been increasingly influenced by

E-mail address: joseramon.saura@urjc.es

https://doi.org/10.1016/j.jik.2020.08.001

research areas such as Information Sciences (IS) or Computer Sciences (CS), as well as by all other areas of research that facilitate collecting, ordering, and managing data (Provost & Fawcett, 2013).

Until now, the key tasks of DS have included improving the storage capacity of company data, performing market research and consumer segmentation, or extracting key information regarding company problems (Loebbecke & Picot, 2015). However, DS is a broad ecosystem that encompasses different pattern identification strategies, models of analysis, performance indicators, statistical variables, and technicalities skills linked to a great technological expertise (Leeflang, Verhoef, Dahlström, & Freundt, 2014). However, several studies have highlighted that there is a skills gap in this industry (e.g., Ghotbifar, Marjani, & Ramazani, 2017; Royle & Laing, 2014). Specifically, both marketers without expertise in IS, CS, or DS and non-technical researchers who lack the knowledge of data management have faced the challenge of acquiring such knowledge and skills and using them not only technically, but also strategically and operationally. These challenges and the ways to overcome them are the motivation of the present study.

In order to analyze the impact of the increase in companies' use of DS on DM, the present study performs an in-depth review of (i) methods of analysis, (ii) uses, and (iii) performance metrics based on DS applied in the scientific literature to DM strategies. The aforementioned three aspects will be studied, explained, and analyzed from the marketer's point of view, rather than from that of a data scientist. By reviewing the main concepts of DS framework applied to DM, the present study will allow marketers and non-technical researchers to better understand the main applications of DS to DM, as well as to become more aware of the importance of each such application.

Considering the scarcity of previous research on the relationship between DS with DM, the present review both bridges a gap in the literature and offers directions of further research in this area. The results of this review will allow marketers and non-technical researchers to better understand how the DS ecosystem applied to DM works, and what the key points of and alternatives to applying these techniques to DM are.

Therefore, the present study addressed the following two research questions:

RQ1: What are the main methods of analysis, uses, and performance metrics of Data Sciences applied in Digital Marketing?

RQ2: What are the areas of further research on the use of Data Science in Digital Marketing?

To address these two questions, a Systematic Literature Review (SLR) is performed based on the publications available from several scientific databases, such as ACM Digital Library, AIS Electronic Library, IEEE Explore ScienceDirect, and Web of Sciences.

The remainder of this paper is structured as follows. The following section presents the theoretical framework of DS. In the second part, Theoretical background, the main concepts of the DS framework are presented. Sections "Methodology development" and "Analysis of results" present the methodology and report the results, with a particular focus on the development of descriptors that define the key aspects of DS in DM. Finally, the last section draws conclusions and outlines directions of further research.

Data science: theoretical framework

Overall, the major goal of DS is to extract knowledge from data analysis to answer specific research questions (Kelleher & Tierney, 2018). By analyzing the data, DS techniques make it possible to extract patterns from databases to explain a problem or to formulate hypotheses. In DS, a key idea is that the patterns identified in the data are (i) non-obvious and (ii) useful for companies (Berry & Linoff, 2004).

Table 1
Types of patterns and descriptions in Data Sciences.

Type of patterns	Patterns description
Clustering	Known as segmentation in business, clustering consists of the identification of behaviors, tastes, or habits that identify the same group of consumers among other uses
Association rule mining (ARM)	ARM consists of identifying patterns of products purchased at the same time
Outlier detection/anomaly	Identification of unusual or uncommon patterns that could be fraudulent, such as in financial or insurance sectors
Prediction patterns (PP)	A pattern that predicts a missing value of an attribute (product, service, etc.) instead of predicting the future

In this relation, it is important to note that, in terms of detecting patterns, humans can identify a maximum of three attributes, or characteristics of an item (product, services, community, etc.). These attributes are also known as features or variables (Saura & Bennett, 2019). However, with DS patterns, hundreds and thousands of attributes (variables) can be simultaneously identified (Berry & Linoff, 2004).

Patterns identified with DS techniques help to obtain actionable insights, i.e. what researchers or data scientists want to extract from the identified patterns (Kelleher, Mac Namee, & D'arcy, 2015). Therefore, the term 'insight' is this context refers to the capacity of patterns to provide meaningful information that can help to solve the problem at stake. The word 'actionable' here means that insights extracted from patterns can in some way be used by the company (Davenport, 2014). In DS, there are different types of patterns that can be applied to the DM industry (see Table 1).

Depending on the company's goals when developing DM, different types of patterns can be used to improve these strategies, as well as to enhance the company's ability to understand and structure the main attributes, features, or variables extracted from companies' databases (Berry & Linoff, 2004). In this sense, the type of collected data is also important, from the company's perspective, as the developed strategies that target digital platforms and social networks should be data-driven (Shareef, Kapoor, Mukerji, Dwivedi, & Dwivedi, 2020).

Large volumes of data are referred to as Big Data (BD) (Gandomi and Haider, 2015). Big Data are characterized by the three Vs: (i) *volume*, i.e. excessive amounts of data, (ii) *variety* of data types, and (iii) *velocity* at which the data must be processed (Kelleher & Tierney, 2018). The foundation of BD as they exist today was laid down by Codd's (1971) Relational Data Model (RDM) that allowed for collecting and storing information, as well as making direct queries on information in databases. This advance removed the concern of the physical location of a database. This was an important milestone in the DS industry—previously, databases were in separate physical storage (Dwork & Roth, 2014).

Codd (1971) also laid the foundations of Structured Query Language (SQL), the current standard for querying databases. The latest developments in data storage have led to the generation of new databases known as NoSQL databases. NoSQL databases store variable data and their attributes with languages such as JavaScript Object Notation (JSON). JSON weighs less, has a higher processing speed, and is self-describing, rather than based on table-based relational model like SQL. Today, data warehouses (DW) are more easily available for analysis, measurement, and control (Janssen et al., 2020).

The origin and sources of the data are also important in DS. Depending on the origin of the data, different types of DS-based approaches are available for each type of analysis. Table 2 shows

Table 2
Types of data sources and descriptions in Data Sciences applied to Digital Marketing.

Type of data	Data description
Transactional data	Information regarding sales, invoices, receipts, shipments, payments, insurance, rentals, etc
Non-transactional data	Demographic, psychographic, behavioral, lifestyle data, etc
Operational data	Data on strategies and actions related to logistics and business operations
Online data (Sources)	User Generated Content (UGC), emails, photos, tweets, likes, shares, websites, web searches, videos, online purchases, music, etc

Table 3
Datasets indicators in Data Sciences.

Dataset	Indicator description
Variable	Denotes an individual abstraction in the dataset
Attribute	Attributes are generally composed of numerical, ordinal, and nominal values. Attributes are characteristics of a variable
Feature	Features respond to the characteristics of variables such as attributes
Entity	Entity is normally defined with a number of attributes
Instance	In DS terms and indicators like instance, the terms example, entity, object, case, individual or record could be used in the DS literature to refer to a row

the main data sources managed by companies that work with DS in DM.

In DS, databases are made up of different variables or indicators. These databases are known as “datasets” or “data records” (thereafter, the term “dataset” will be used). Each of the variables contained in the datasets denotes a specific characteristic (see Table 3). Datasets can contain (i) structured or (ii) unstructured data. Structured data can be stored in tables, with each table having the same structure or attributes. By contrast, unstructured data have their own internal structure and, consequently, attributes can be organized in different ways in each table.

To provide an idea of the organization of datasets used by marketers applying DS, Table 3 presents the main characteristics of the indicators typically used in such datasets.

Attributes sometimes come from raw data. Raw data include different types of data that can offer insights to solve a problem. According to Kitchin (2014), raw data can be divided into (i) capture data and (ii) exhaust data. Capture data include measurements and observations that have been designed to collect data. By contrast, exhaust data do not include such measurements and observations and must be structured based on the problem to be solved. In previous research, case studies on metadata—i.e., data that contain files uploaded to the Internet—are analyzed by researchers as exhaust data (Janssen et al., 2020).

For dataset analysis, DS rely on the models based on Machine Learning (ML). The core of modern DS, ML provides algorithms to automatically analyze large datasets. These models can be trained by researchers (also non-technical researchers) or marketers to extract actionable insights and identify patterns. A wide array of algorithms is available. These algorithms can be used and trained by connecting to companies or researchers' databases (Dwork & Roth, 2014).

ML has evolved into what is known as Deep Learning (DL), a technology that has allowed us to change how computers process language and images. DL consists of a set of neural network models with multiple layers and units on the same network. DL is the newest form of ML; however, it is not the only one used in DS (Lies, 2019). There are different other approaches to the data using ML (see Table 4 for a summary).

Table 4
Main machine learning models.

Type	Description
Ensemble models (EM)	EM are used to make predictions using a model source where each model votes on each query
Deep-learning neural networks	Deep-learning neural networks have different layers of networks that discover patterns and can be trained. Models can also learn these patterns from complex datasets and can be applied to others later by following the learned criteria
Machine Vision (MV)	Within the deep-learning neural networks, MV allows for a visual identification and recognition of objects, people, products, etc
Natural-Lenguage Processing (NLP)	Based on text analysis and predefined patterns, NLP works with language and text to identify insights that explain patterns non-identifiable by humans

Table 5
Machine learning approaches applied to Digital Marketing.

Type	Description
Supervised learning (SL)	SL the action of the ML that allows an algorithm to map an input to an-output normally known as input–output pairs
Unsupervised learning (SL)	SL is the action of the ML that allows the function of identify for previously undetected patterns in a dataset with no pre-existing labels
Support Vector Machines (SVM)	Support vector is a one-class support vector machine algorithm that classifies data into simple units and examines how similar instances are in a dataset

Next, within the ML area, there are two main types of analysis approaches: (i) Supervised Learning (SL) and (ii) Unsupervised Learning (UL) (see Table 5). SL involves training a set of samples, including pieces of text, User Generated Content (UGC), such as tweets or Facebook posts, feelings about a product, and so forth. All these samples can be used to train an algorithm. Once the expected success rate (accuracy) of the algorithm is achieved, the algorithm that works with ML can automatically perform the analysis automatically.

Unlike SL, Unsupervised Learning automatically identifies patterns that have not yet been detected in the data. In this case, human supervision is minimal. The algorithms that work with ML are called as Support Vector Machines (SVM), known as one-class classifiers. An SVM examines the dataset to identify the main characteristics and similar behaviors of the instances that make up the database and can be trained on a recurring basis. SVM will classify the dissimilar values of the instances so that the investigator can study the identified anomalies.

Data analysis processes should be framed using relevant concepts. Accordingly, in the DS field, the concepts of Data Mining (DMI) and Knowledge Discovery (KD) have been introduced. At present, these two concepts (DMI and KW) are used indiscriminately by researchers to refer to dataset analysis strategies. The concept of DMI, proposed by Lovell and Michael (1983), initially emerged in the business world to make sense of the datasets developed in data warehouses. Accordingly, today, the concept of DMI is more used in the world of business and marketing to refer to processes of discovery and identification of patterns in datasets. By contrast, KD stems from the concept of Knowledge Discovery in Databases (KDD) (Shapiro, 1989), a concept that is now more extensively used in the scientific world. It is a more technical concept to refer to processes of discovery and identification of patterns in datasets (Rudder, 2014).

Table 6
Search terms used in the SLR.

Search terms			Data bases	Fields
Data sciences	AND	digital marketing	ACM Digital Library	Title
OR data mining*		OR online marketing*	AIS Electronic Library	Abstract
OR knowledge* discovery		OR internet marketing*	IEEE Explore	Keywords
			ScienceDirect	
			Web of Sciences	

* These terms were only used when the search of the terms "Data Sciences and Digital Marketing" did not yield the expected results.

Methodology development

In this study, in order to address the research questions formulated in the previous section, the methodology of a Systematic Literature Review (SLR) was used. SLP is defined as a method that enables tackling "*an emerging issue that would benefit from exposure to potential theoretical foundations*" (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018; Webster & Watson, 2002). As discussed previously, the emerging area of DS applied to the DM sector will benefit from a logical conceptualization of the application of DS to this digital environment.

For the development of SLR, the following three steps are usually used (Stieglitz et al., 2018). First, the theoretical foundations of a framework are presented for the classification of the main DS concepts used in the literature. In doing so, we follow Bem (1995) who argued that a coherent review requires a conceptual coherent structuring of the topic itself. Therefore, the present review focused on the analysis methods, uses, and performance metrics for the use of DS in DM. The main contributions of relevant studies were identified and categorized in terms of their priority for the theoretical framework.

In the second step, according to Stieglitz et al. (2018), the literature is systematically examined to identify similarities and details of the DM sector in DS. This step is used to inductively synthesize prior research and group basic concepts and definitions (Devece, Ribeiro-Soriano, & Palacios-Marqués, 2019). In the third step, the main findings of the analysis of DS in DM in the literature are considered, highlighting the main uses, applications, indicators and techniques, as well as outlining directions of further research in this field.

In the present study, we followed the guidelines proposed by vom Brocke et al. (2009, 2015) and Stieglitz et al. (2018). First, the predefined and selected terms were searched in the databases regarding title, abstract, and keywords. The irrelevant results were eliminated. The search terms were chosen to identify the main uses, applications, indicators and techniques, as well as the future of DS in DM according to the theoretical framework (see Table 6). The results were classified by categories and filters related to CS, IS, DM, marketing and business. Furthermore, only original articles and reviews were analyzed. Proceedings, books, chapters or magazines were not included in the SLR process.

This study is database-oriented and took into account all articles published and indexed in the following scientific databases: ACM Digital Library, AIS Electronic Library, IEEE Explore, ScienceDirect, and Web of Science. The detailed list of search is shown in Table 6.

Second, to identify the potential of the found articles, titles, abstracts, and keywords were read in detail. For our review, relevant research studies were those that identified the main analysis methods, uses, performance metrics and future of DS research in DM; therefore, the type of analysis and methodologies used in the studies were not taken into account when selecting the articles.

Table 7
SLR results.

Database	Number of results	Number of relevant results
ACM Digital Library	136	13
AIS Electronic Library	5	2
IEEE Explore	79	10
ScienceDirect	34	9
Web of Sciences	105	15
Total	354	49

Table 8
Number of article classifications by results.

Article classification	Data Sciences	Digital Marketing	Mixed Articles
Classification	16	13	20
Methods	12	4	11
Uses	6	6	11
Performance metrics	4	11	11
Future topics	4	5	9

Thirdly, selected articles were categorized as relevant following the definitions, applications, and theoretical concepts regarding the importance of applying DS techniques in DM. Therefore, the articles that contained inadequate terms or were not conclusive, did not contain the search terms, had no relation to the research topic, offered no quality evaluation, or contained no description and specification of terms were removed (see Fig. 1).

The steps of the development of the SLR undertaken in the present study, performed following vom Brocke et al. (2015) and Stieglitz et al. (2018) and enriched by the SLR presented by the PRISMA diagram (Moher, Liberati, Tetzlaff, Altman, & T, 2009), are shown in Fig. 1.

Table 7 shows the total number of articles identified based on each of the objectives proposed in the SLR process.

While the very nature of the SLR indicates that the results from the databases should be related to both DS and DM, the articles were classified from the point of view of the factors analyzed. That is, if some articles were classified as having the main focus of DS, the analyzed factors were justified from the DS point of view. On the other hand, for those articles classified as focusing on DM, the concepts analyzed were from the point of view of DM and its influence on DS. Furthermore, there were articles analyzed from a broader point of view as they focused on methods, uses, performance metrics and future topics for both categories. In addition, articles that presented different categories (methods, uses, performance metrics and future topics) within the same article were also analyzed (see Table 8).

Table 9 summarizes all studies included in the review, with the specifications of authors, journals, research category, and content (specifically, main topic and main focus).

Analysis of results

DS provide different perspectives and approaches to statistical data analysis. Statistics is a set of rules to quantitatively analyze any type of data. However, with the evolution of mathematics and the development of DS, statistical learning has been defined as a theoretical framework that works with ML from the point of view of DS (Tsapatsoulis & Djouvas, 2019). Therefore, the objective of the methods used in DS applying statistical learning is to perform (i) functional analysis, (ii) exploratory analysis, and (iii) prediction of results based on the analyzed data. Table 10 provides a summary of the main methods identified in and applied to the DM ecosystem.

After the analysis of the main methods applied to the study of DM using the DS techniques, we identified the uses and applications

Table 9
Relevant studies found in the Systematic Literature Review.

Authors	Journal	Category	Main topic		Main focus			
			Data Sciences	Digital Marketing	Methods	Uses	Perform. Metric	Future Topics
Alrifai (2017)	The Journal of Computing Sciences in Colleges	Computer Sciences	•		•	•	•	
Arias, Arratia, and Xuriguera (2014)	ACM Transactions on Intelligent Systems and Technology	Intelli. Sys. & Tech.	•	•	•		•	
Ballestar, Grau-Carles, and Sainz (2018)	Journal of Business Research	Business		•		•	•	
Beaput, Banik and Joshi (2019)	The Journal of Computing Sciences in Colleges	Computer Sciences	•					•
Cassavia, Masciari, Pulice, and Sacca (2017)	ACM Transactions on Interactive Intelligent Systems	Interactive Intelli. Sys.	•	•	•		•	
Cheng and Wang (2018)	Journal of Business Economics and Management	Business, Economics		•	•		•	•
Dadzie, Sibarani, Novalija, and Scerri (2018)	Journal of Web Semantics	Information Systems	•		•			
De Caigny, Coussement, and De Bock (2020)	Decision Support Systems	Computer Sciences	•	•	•	•		•
Debortoli, Müller, and vom Brocke (2014)	Business & Information Systems Engineering	Business, Info. Systems	•	•		•	•	•
Dwivedi et al. (2019)	International Journal of Information Management	Information Sciences	•			•		•
Fan et al. (2015)	Proceedings of the VLDB Endowment	Computer Sciences	•		•	•		
Guinan, Parise, and Langowitz (2019)	Business Horizons	Business		•		•		•
Haq, Li, and Hassan (2019)	IEEE Access	Computer Sciences	•		•	•		
Hodge et al. (2018)	IEEE Transactions on Games	Computer Sciences	•			•		•
Jacobson, Gruzd, and Hernández-García (2020)	Journal of Retailing and Consumer Services	Business	•	•		•		•
Jiao, Wang, Feng, and Niyato (2018)	IEEE Internet of Things Journal	Computer Sciences	•		•			•
Lagrée, Cappé, Cautis, and Maniu (2018)	ACM Transactions on Knowledge Discovery from Data	Information Science		•		•	•	
Lau et al. (2012)	ACM Transactions on Management Information Systems	Information Systems	•	•	•	•		
Liu et al. (2016)	ACM Transactions on Knowledge Discovery from Data	Information Sciences		•	•	•	•	
Liu, Gu, Ko, and Liu (2018)	IEEE transactions on cybernetics	Automa. & Control Syst.	•	•		•	•	
Miklosik, Kuchta, Evans, and Zak (2019)	IEEE Access	Computer Sciences	•	•		•	•	
Milovanović, Bogdanović, Labus, Barać, and Despotović-Zrakić (2019)	Future Generation Computer Systems	Computer Sciences	•		•	•		
Naqvi, Awais, Saeed, and Ashraf (2018)	Inter. J. of Advanced Computer Science and Applications	Com. Sc.,The. & Metho.	•		•		•	
Papoutsoglou, Ampatzoglou, Mittas, and Angelis (2019)	IEEE Access	Computer Sciences	•		•			

Table 9 (*Continued*)

Authors	Journal	Category	Main topic		Main focus			
			Data Sciences	Digital Marketing	Methods	Uses	Perform. Metric	Future Topics
Poddar, Banerjee, and Sridhar (2019)	Journal of Business Research	Business		•		•	•	
Roundy and Asllani (2018)	Journal of Entrepreneurship, Management and Innovation	Business, Economics	•	•		•	•	•
Ruggieri, Pedreschi, and Turini (2010)	ACM Transactions on Knowledge Discovery from Data	Information Sciences	•		•		•	
Saleemi, Anjum, and Rehman (2017)	IEEE Access	Computer Sciences	•		•	•		•
Salminen, Yoganathan, Corporan, Jansen, and Jung (2019)	Journal of Business Research	Business	•	•	•		•	•
Sang and Xu (2011)	ACM Trans. Mult. Comp., Commu., and Applications	Computer Science	•	•	•		•	•
Saura and Bennett (2019)	Symmetry-Basel	Multidisciplinary	•	•	•	•	•	
Saura, Herráez, and Reyes-Menendez (2019)	IEEE Access	Computer Sciences	•	•	•		•	
Spiekermann, Korunovska, and Langheinrich (2018)	Proceedings of the IEEE	Engineering	•					•
Troisi, Maione, Grimaldi, and Loia (2019)	Industrial Marketing Management	Business, Management		•			•	•
Tsapatsoulis and Djouvas (2019)	Frontiers in Robotics and AI	Robotics		•	•		•	
Wu and Holsapple (2013)	ACM Transactions on Knowledge Discovery from Data	Information Sciences		•		•	•	•
Wymbs (2016)	Journal of Information Systems Education	Information Systems		•		•	•	
Yeo, Hwang, Kim, Koh, and Lipka (2018)	IEEE Transactions on Knowledge and Data Engineering	Information Sciences	•	•	•	•		
Zannettou, Sirivianos, Blackburn, and Kourtellis (2019)	Journal of Data and Information Quality	Data and Infor. Quality	•	•		•	•	
Zhu, Peng, Chen, Zheng, and Zhou (2016)	ACM Transactions on Knowledge Discovery from Data	Information Sciences		•		•	•	•
Wang, Zhang, and Yuan (2017)	Foundations and Trends® in Information Retrieval	Computer Sciences	•	•			•	•
Song et al. (2014)	Personal and ubiquitous computing	Computer Sciences	•	•	•	•		
Chang, Kauffman, and Kwon (2014)	Decision Support Systems	Computer Sciences	•					•
Kaiser and Bodendorf (2012)	Internet Research: Electronic Netw. Applic. and Policy	Business & Info. Sci.		•		•	•	
Archak et al. (2011)	Management science	Management	•	•		•		•
Liao, Chen, Hsieh, and Hsiao (2009)	Expert Systems with Application	Computer Sciences	•	•	•		•	•
Su, Chen, and Sha (2007)	International Journal of Technology Management	Management		•	•	•		•
Su, Chen, and Sha (2006)	Technovation	Engineering & Managem.	•		•		•	
Kwan, Fong, and Wong (2005)	Decision Support Systems	Computer Sciences	•	•		•	•	•

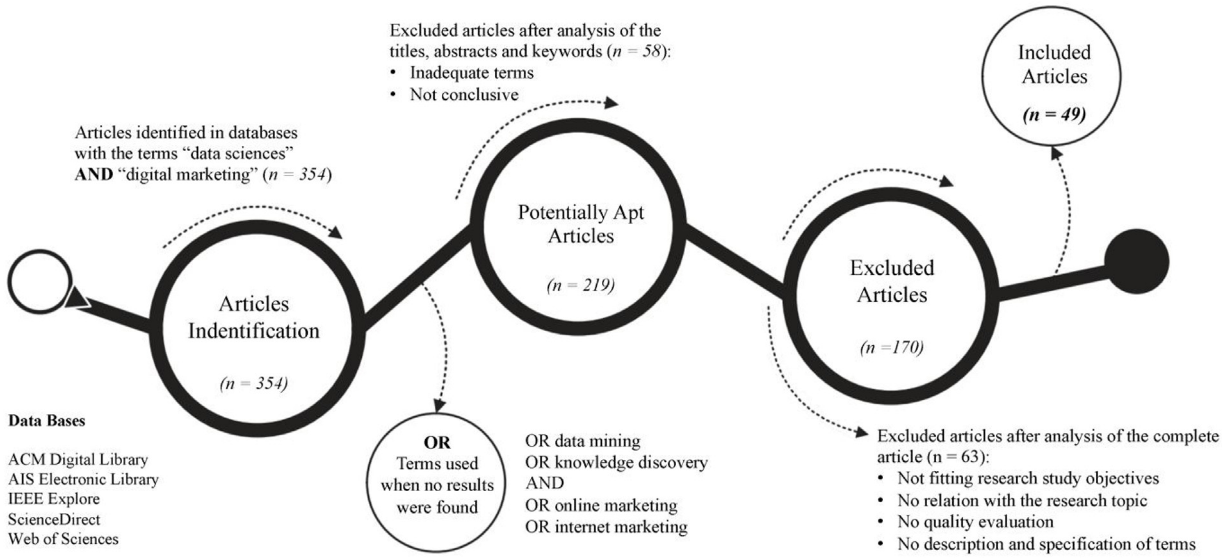


Fig. 1. PRISMA flow diagram of SLR.

Table 10
Type of methods used in Data Sciences as applied to Digital Marketing.

Type of analysis	Description
Descriptive statistics	Descriptive statistics is used to quantitatively summarize features from a collection of information or a dataset. It includes the measurement of central tendency, arithmetic mean, or measures such as variance or the range
Bayes' rule	Bayer's rule is used for the description of probabilities of events. It is based on knowledge about the conditions that could have caused a specific event
Method of least squares	The method of least squares makes it possible to find the best theoretical model made up of variables or constructs that fit a dataset and allows for a quantitative validation
Linear regression	Linear regression is used to model the relationship between a scale and an exploratory variable. When there is more than one analysis variable within the linear regression, it is called multiple linear regression
Logistic regression	Logistic regression is a regression analysis used to predict a categorical variable. A categorical variable is a variable that can be subdivided into more categories. It is used as a general rule to model the probability of an event and the factors that compose it
Artificial neural network	Artificial neural networks are self-learning systems made up of interconnected neurons of nodes that have an input and an output. They are used to find or detect solutions or features that are otherwise difficult to identify using standard programming
Multivariate analysis	Multivariate analysis is an analysis model focused on the multiple analysis of data collected from more than one dependent variable. Multivariate analysis involves the analysis of variables and correlations among them
Maximum likelihood estimate	Maximum likelihood estimate is a method for estimating statistical parameters in an observation. Its results show the highest probability for the estimation of the analyzed parameters
Discriminatory analysis	Discriminatory analysis is known as classification or pattern-recognition (nearest-neighbor models). It automatically recognizes patterns or regularities in a dataset and is widely used in research focused on DM or KDD
Information theory	Information theory is used to analyze and study the quantification and communication of information to find the limits of communication processing
Artificial Intelligence	Artificial Intelligence is focused on the simulation of human intelligence by machines. In the present study, Artificial Intelligence refers to using machines automatically focused on machine learning and problem solving

pursued by the DS in the tactical strategies and developments of the DM (see Table 11).

In the DM ecosystem, one of the main challenges is controlling and defining the success of a DS strategy. To this end, marketers and researches should choose and understand the main performance metrics for the measurement of the models and methodologies used. The results of our SLR analysis highlighted the following indicators to measure the success of these SD strategies applied to the DM sector (see Table 12).

Fig. 2 shows the main topics DM research using DS. These research areas were selected taking into account the recommendations of the studies analyzed and their discussions of further research in the field.

In what follows, each of the topics and the influence that these topics may have on the development of DM strategies using DS are explained in detail.

Medical data and eHealth strategies: Analysis of users' medical data can help to find trends and thus facilitate the creation of new vaccines, fighting diseases that can cause an uncontrolled epidemic (such as the coronavirus known as COVID-19), and predicting possible deaths. Marketing must promote companies' use of such strategies to collect data for further analysis.

Smart cities & governance: Efficient management of energy resources, as well as sustainable and intelligent construction and development are based on automation and artificial intelligence of large structures (Ismagilova, Hughes, Dwivedi, & Raman, 2019; Janssen, Luthra, Mangla, Rana, & Dwivedi, 2019). Social or responsible marketing, also known as corporate social responsibility (CSR), is driven by DM-based communications through digital platforms and social media channels (Orlandi, Ricciardi, Rossignoli, & De Marco, 2019).

Internet of Things (IoT): IoT refers to management and collection of daily use data from connected devices. This also includes order and identification of new features that help personalize and offer new products and services and to create new needs (Brous, Janssen, & Herder, 2020). DM adapts to the mobile environment

Table 11
Mean uses of Data Sciences in Digital Marketing strategies.

Type	Description
Analyzing User Generated Content (UGC)	Analysis of the content publicly published by users on social networks and digital platforms. Some examples include tweets, reviews, reviews, comments, shares, likes, meta data, etc
Optimize customers' preferences	Optimization of user preferences in products or services customizable based on the analysis of large databases containing information about customer preferences
Tracking customer behavior online	Tracking and predicting consumer behavior on digital channels. In this way, companies can anticipate their marketing and advertising actions and better understand users
Tracking social media commentary/Interactions	Tracking user-company responses and interactions in digital ecosystems. Based on the results, marketers can transmit more relevant messages
Optimize stock levels in ecommerce websites	Optimization of the stock in e-commerce websites for better management of both the information and the products and services to anticipate sales and request products from suppliers in advance
Analysis of online sales data	Sales analysis to find patterns that explain trends, current demand and customers' interest in specific products
Introducing new products	Market research to find implausible patterns that can offer new products to sectors saturated with offers
Analyzing social media trends	Analysis of the trends in social media to avoid, for example, crisis of reputation of companies or social movements that require attention by the company. Also known as Social Listening
Analyzing product recommendations and reviews	Analysis of consumer reviews and recommendations to identify key points and characteristics of companies' products, services, or customer service
Personalizing customer's online experience	Improving customer shopping experiences by offering personalized experiences based on information from previous users who visited a website
Building recommender systems	Creation of automatic recommendation systems that establish priorities and predictions of purchase success for certain customers based on the analysis of their purchase history
Measure and predict clicks online (Social and Paid Ads)	Prediction of user clicks on web pages and social networks. Based on this information, companies can improve the profitability of their ads and increase the visibility of impressions and clicks, also known as Click Through Rate (CTR)
Measure and predict user's behavior	Predicting how users will behave to offer or avoid problems before a purchase, e.g. sending emails in email marketing campaigns to offer additional products and thus increase profitability
Improve User Experience (UX)	Improving user experience in terms of graphic and visual design of web pages, mobile applications, etc. focused on the "user-friendly" movement
Analyze real-time data	Facilitation of direct decisions regarding certain situations of sales, visibility, or dissemination of information
Identify online communities	Identification of online communities and their influential and leaders. Analysis of the user considerations and loyalty to opinion leaders and their messages
Identificar fake news y contenidos falsos	Identification of false content about companies or shared advertising messages; alternative terms are "fakenews" or "deepfakes"

Table 12
Mean performance metrics to measure the success of DS approaches in DM.

Indicator	Description
Reliability	Reliability is the metric of accuracy and exhaustiveness of the processing of a dataset linked to its use
Accuracy	Accuracy, also referred to as precision, shows the quality and accuracy of hit of a model or method
Precision	Precision, also known as Positive Predictive Value (PPV), is a metric that measures the relevance and success of the approach of a method within the database where DS has been applied
Validity	Validity is the measure that indicates whether the data support the results and conclusions obtained in a truthful and accurate way.
Consistency	Consistency evaluates whether the values represented in data from one dataset are consistent with the values represented in the data from another dataset at the same point in time
Recall	Recall, also called sensitivity, generally refers to the number of correct results divided by the number of discarded values
Sensitivity	Sensitivity, sometimes referred to as probability of detection or positive rate, measures the proportion of positive values or points identified in a dataset
Specificity	Specificity is the metric that evaluates the prediction characteristics of true negatives in the variables within a category in a dataset
Prevalence	Prevalence refers to the proportion of population with specific common characteristics during a specific period of time

with mobile-friendly design initiatives and strategies that fosu on connected devices.

Data privacy and management: Mass data management: This includes rights, access, and legitimate profitability of large public databases (Nissenbaum, 2009). One of the functions of DM is to raise consumer awareness about how companies will make use of their data (emails, phones, demographics, and so forth) (Lee & Trimi, 2018).

People: movement, organization and personalization: This includes analyzing the movement and organization of people through the analysis of large databases of citizens or vehicle purchases (Aladwani & Dwivedi, 2018; Höflinger, Nagel, & Sandner, 2018). The DM has the challenge of personalizing massive messages and, using the DS methods, identifying specific habits according to the type of people and their demographic and psychographic characteristics to increase the ROI of digital campaigns (Palacios-Marqués, García, Sánchez, & Mari, 2019).

Development of new Machine Learning models: This includes new Machine Learning models that companies can train and apply in their projects. There is a growing need for the creation of models focused on solving specific problems. These models should be created, trained, and debugged for a specific purpose. Other tasks include designing user-friendly algorithms and ML models to break the technical barrier between marketers and data specialists (Caseiro & Coelho, 2019).

Operational CRM and data management: This includes the creation of automatic company information management systems that can identify better unsuspected patterns and extract actionable insights to help company to manage their information in real time and to enable marketing experts to take better decisions (Ricciardi, Zardini, & Rossignoli, 2018).

Sustainable strategies based on data: This includes the study of the sources of data resources and management of globalization processes to increase sustainable strategies and actions based on data analysis. Relevant research areas in this field include social marketing or green marketing (Archak, Ghose, & Ipeirotis, 2011).

Social media listening: This includes automated research on important trends in social networks and messages released by opinion leaders, as well as exploring the responses of communities to massive messages in the face of crises, epidemics, as well

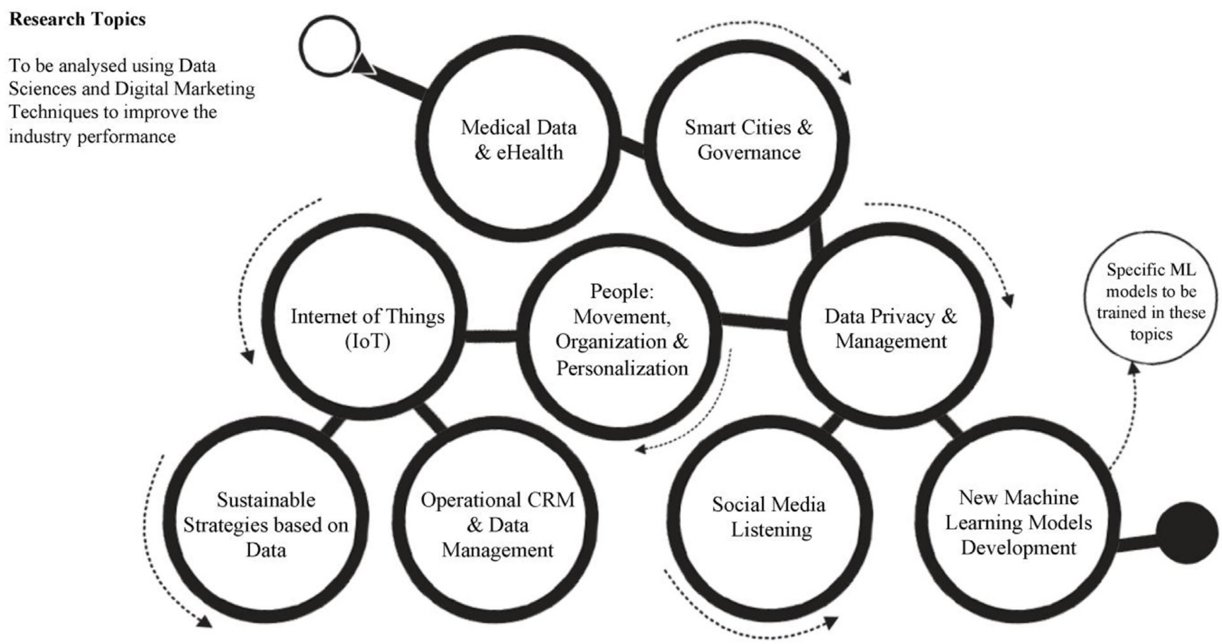


Fig. 2. Main topics of research in Digital Marketing using Data Sciences.

as environmental or social movements (Reyes-Menendez, Saura, & Stephen, 2020). DM should understand how these communities are organized and take adequate action with persuasive and responsible messages.

Conclusions

In this review article, we have defined the main concepts, methods, and performance metrics used in DS throughout the last two decades and their applications in DM. We have provided a structured account of the main concepts that marketers should take into account when considering a DM strategy based on data intelligence. Pertinent methods used in DS to extract actionable insights from large amounts of data have also been identified. We have also outlined major performances metrics used to measure the DS performance in the DM environment. These results respond to first research question addressed in the present study (*What are the main methods of analysis, uses, and performance metrics of Data Sciences applied in Digital Marketing?*).

Today, companies are involved in an increasingly data-driven ecosystem. Accordingly, the number of ML-based user-friendly applications that companies, marketers, and non-technical researchers can use has considerably increased. However, the understanding by marketers and marketing researchers of the main notions of DS is essential to be efficient and lasting over time, as the lack of such understanding has already become a skill problem (Ghotbifar et al., 2017; Royle & Laing, 2014).

Specifically, businesses have been reported to waste a lot of time organizing, cleaning, and structuring the databases of their users and customers (Kelleher & Tierney, 2018). In this context, the use of relevant indicators and performance metrics will help companies, marketers, and non-technical researchers in the marketing area to conduct better research and to more efficiently measure the time they spend analyzing and structuring their databases.

With regard to our second research question ("*What are the areas of further research on the use of Data Science in Digital Marketing?*"), in our results, we have identified a total of 9 topics for future research on DS in the DM ecosystem. Undoubtedly, the application of new specific ML models to each of these topics will define the future of the sector in terms of the effectiveness of its data-driven strategies.

As for the theoretical implications, the present review has identified a total of 11 methods, 17 uses, 9 performance metrics, and 9 research topics that can be used by researchers as the starting points of their research focused on the use of DS in strategies of DM. Regarding the methods, when considering their DM research using DS analysis, non-technical researchers can take into account which of these models best fit the objectives of their research based on the definitions presented. Furthermore, for the elaboration of new studies, researchers in the DM sector can use the nine identified topics to formulate new hypotheses or to find new research questions that need be addressed.

Furthermore, the present review offers important practical implications for the industry. Today, companies are increasingly developing data-driven strategies. Therefore, the best use of these strategies requires an in-depth understanding of all necessary steps. The results of the present review can be meaningfully used to familiarize companies' experts with the main DS indicators and metrics in the DM ecosystem.

Consequently, companies can use the results of the present review as the starting point in the elaboration of new DM strategies. Companies can consider implementing any of the 17 identified uses of DS in DM to obtain actionable insights from their datasets. In addition, in terms of performance metrics, companies can take the definitions and descriptions provided in this review to make reports and present their contingency and control plans, as well as to measure the success of their digital campaigns.

The limitations of this study include the number of databases analyzed and the criteria used to collect the articles from the databases. In addition, the selection of the articles and their classification could have biased the final results. Further research should focus on the topics similar to the 9 topics identified in the present review, In addition, the improvement of DS processes in the DM ecosystem, which has not been identified in this review, should also be considered.

References

Aladwani, A. M., & Dwivedi, Y. K. (2018). Towards a theory of SocioCitizenry: Quality anticipation, trust configuration, and approved adaptation of governmental social media. *International Journal of Information Management*, *43*, 261–272.

Alrifai, R. (2017). A data mining approach to evaluate stock-picking strategies. *Journal of Computing Sciences in Colleges*, *32*(5), 148–155.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485–1509.

Arias, M., Arratia, A., & Xuriguera, R. (2014). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 1–24.

Avery, J., Steenburgh, T. J., Deighton, J., & Caravella, M. (2012). Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time. *Journal of Marketing*, *76*(3), 96–111.

Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2018). Customer segmentation in e-commerce: Applications to the cashback business model. *Journal of Business Research*, *88*, 407–414.

Beaput, A., Banik, S., & Joshi, D. (2019). Ranking privacy of the users in the cyberspace. *The Journal of Computing Sciences in Colleges*, *35*(4), 109–114.

Bem, D. J. (1995). Writing a review article for Psychological Bulletin. *Psychological Bulletin*, *118*(2), 172–177. http://dx.doi.org/10.1037/0033-2909.118.2.172

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. Hoboken, New Jersey: John Wiley & Sons.

Braverman, S. (2015). Global review of data-driven marketing and advertising. *Journal of Direct Data and Digital Marketing Practice*, *16*(3), 181–183.

Brous, P., Janssen, M., & Herder, P. (2020). The dual effects of the Internet of Things (IoT): A systematic review of the benefits and risks of IoT adoption by organizations. *International Journal of Information Management*, *51*, 101952.

Caseiro, N., & Coelho, A. (2019). The influence of Business Intelligence capacity, network learning and innovativeness on startups performance. *Journal of Innovation & Knowledge*, *4*(3), 139–145. http://dx.doi.org/10.1016/j.jik.2018.03.009

Cassavia, N., Masciari, E., Pulice, C., & Sacca, D. (2017). Discovering user behavioral features to enhance information search on big data. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(2), 1–33.

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, *63*, 67–80.

Cheng, F. C., & Wang, Y. S. (2018). The do not track mechanism for digital footprint privacy protection in marketing applications. *Journal of Business Economics and Management*, *19*(2), 253–267.

Codd, E. F. (1971). A data base sublanguage founded on the relational calculus. In *Proceedings of the 1971 ACM SIGFIDET (now SIGMOD) workshop on data description, access and control* (pp. 35–68). November.

Dadzie, A. S., Sibarani, E. M., Novalija, I., & Scerri, S. (2018). Structuring visual exploratory analysis of skill demand. *Journal of Web Semantics*, *49*, 51–70.

Davenport, T. (2014). *Big data at work: dispelling the myths, uncovering the opportunities*. Brighton, Massachusetts: Harvard Business Review Press.

De Caigny, A., Coussement, K., & De Bock, K. W. (2020). Leveraging fine-grained transaction data for customer life event predictions. *Decision Support Systems*, *130*, 113232.

Debortoli, S., Müller, O., & vom Brocke, J. (2014). Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, *6*(5), 289–300.

Devece, C., Ribeiro-Soriano, D. E., & Palacios-Marqués, D. (2019). Coopetition as the new trend in inter-firm alliances: literature review and research patterns. *Review of Managerial Science*, *13*(2), 207–226.

Dremel, C., Herterich, M. M., Wulf, J., & Vom Brocke, J. (2020). Actualizing big data analytics affordances: A revelatory case study. *Information & Management*, *57*(1), 103121.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 101994.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

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.

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

Ghotbifar, F., Marjani, M., & Ramazani, A. (2017). Identifying and assessing the factors affecting skill gap in digital marketing in communication industry companies. *Independent Journal of Management & Production*, *8*(1), 1–14.

Guinan, P. J., Parise, S., & Langowitz, N. (2019). Creating an innovative digital project team: Levers to enable digital transformation. *Business Horizons*, *62*(6), 717–727.

Haq, M. I. U., Li, Q., & Hassan, S. (2019). Text mining techniques to capture facts for cloud computing adoption and big data processing. *IEEE Access*, *7*, 162254–162267.

Hodge, V., Sephton, N., Devlin, S., Cowling, P., Goumagias, N., Shao, J., et al. (2018). How the business model of customisable card games influences player engagement. *IEEE Transactions on Games*, *11*(4), 374–385.

Höflinger, P. J., Nagel, C., & Sandner, P. (2018). Reputation for technological innovation: Does it actually cohere with innovative activity? *Journal of Innovation & Knowledge*, *3*(1), 26–39. http://dx.doi.org/10.1016/j.jik.2017.08.002

Ismagilova, E., Hughes, L., Dwivedi, Y. K., & Raman, K. R. (2019). Smart cities: Advances in research – An information systems perspective. *International Journal of Information Management*, *47*, 88–100.

Jacobson, J., Gruzd, A., & Hernández-García, Á. (2020). Social media marketing: Who is watching the watchers? *Journal of Retailing and Consumer Services*, *53*, 53.

Janssen, M., Luthra, S., Mangla, S., Rana, N. P., & Dwivedi, Y. K. (2019). Challenges for adopting and implementing IoT in smart cities. *Internet Research*, *29*(6), 1589–1616.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 101493.

Jiao, Y., Wang, P., Feng, S., & Niyato, D. (2018). Profit maximization mechanism and data management for data analytics services. *IEEE Internet of Things Journal*, *5*(3), 2001–2014.

Kaiser, C., & Bodendorf, F. (2012). Mining consumer dialog in online forums. *Internet Research: Electronic Networking Applications and Policy*, *22*(3), 275–297.

Kannan, P. K. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing*, *34*(1), 22–45.

Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. Cambridge, Massachusetts: MIT Press.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Newbury Park, California: Sage.

Kumar, V., Chattaraman, V., Neghina, C., Skiera, B., Aksoy, L., Buoye, A., et al. (2013). Data-driven services marketing in a connected world. *Journal of Service Management*, *24*(3), 330–352.

Kwan, I. S., Fong, J., & Wong, H. K. (2005). An e-customer behavior model with online analytical mining for internet marketing planning. *Decision Support Systems*, *41*(1), 189–204.

Lagrée, P., Cappé, O., Cautis, B., & Maniu, S. (2018). Algorithms for online influencer marketing. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *13*(1), 1–30.

Lau, R. Y., Liao, S. Y., Kwok, R. C. W., Xu, K., Xia, Y., & Li, Y. (2012). Text mining and probabilistic language modeling for online review spam detection. *ACM Transactions on Management Information Systems (TMIS)*, *2*(4), 1–30.

Lee, S. M., & Trimi, S. (2018). Innovation for creating a smart future. *Journal of Innovation & Knowledge*, *3*(1), 1–8. http://dx.doi.org/10.1016/j.jik.2016.11.001

Leeflang, P. S., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. *European Management Journal*, *32*(1), 1–12.

Liao, S. H., Chen, C. M., Hsieh, C. L., & Hsiao, S. C. (2009). Mining information users' knowledge for one-to-one marketing on information appliance. *Expert Systems with Applications*, *36*(3), 4967–4979.

Lies, J. (2019). Marketing intelligence and big data: digital marketing techniques on their way to becoming social engineering techniques in marketing. *International Journal of Interactive Multimedia & Artificial Intelligence*, *5*(5).

Liu, B., Wu, Y., Gong, N. Z., Wu, J., Xiong, H., & Ester, M. (2016). Structural analysis of user choices for mobile app recommendation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *11*(2), 1–23.

Liu, Y., Gu, Z., Ko, T. H., & Liu, J. (2018). Identifying key opinion leaders in social media via modality-consistent harmonized discriminant embedding. *IEEE Transactions on Cybernetics*, *50*(2), 717–728.

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems*, *24*(3), 149–157.

Lovell, & Michael, C. (1983). "Data Mining". *The Review of Economics and Statistics*, *65*(1), 1–12. http://dx.doi.org/10.2307/1924403

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston, Massachusetts: Houghton Mifflin Harcourt>.

Miklosik, A., Kuchta, M., Evans, N., & Zak, S. (2019). Towards the adoption of machine learning-based analytical tools in digital marketing. *IEEE Access*, *7*, 85705–85718.

Milovanović, S., Bogdanović, Z., Labus, A., Barać, D., & Despotović-Zrakić, M. (2019). An approach to identify user preferences based on social network analysis. *Future Generation Computer Systems*, *93*, 121–129.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and MetaAnalyses: The PRISMA statement. *PLoS Med*, *6*(7), e1000097. http://dx.doi.org/10.1371/journal.pmed1000097

Naqvi, S. M. A., Awais, M., Saeed, M. Y., & Ashraf, M. M. (2018). Opinion mining and thought pattern classification with natural language processing (NLP) tools. *International Journal of Advanced Computer Science and Applications*, *9*(10), 485–493.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Redwood City, California: Stanford University Press.

Orlandi, L. B., Ricciardi, F., Rossignoli, C., & De Marco, M. (2019). Scholarly work in the Internet age: Co-evolving technologies, institutions and workflows. *Journal of Innovation & Knowledge*, *4*(1), 55–61. http://dx.doi.org/10.1016/j.jik.2017.11.001

Palacios-Marqués, D., García, M. G., Sánchez, M. M., & Mari, M. P. A. (2019). Social entrepreneurship and organizational performance: A study of the mediating role of distinctive competencies in marketing. *Journal of Business Research*, *101*, 426–432.

Palacios-Marqués, D., Welsh, D. H., Méndez-Picazo, M. T., Palacios-Marqúes, D., Welsh, D. H. B., & Méndez-Picazo, M. T. (2016). The importance of the activities of innovation and knowledge in the economy: Welcome to the Journal of Innovation and Knowledge. *Journal of Innovation and Knowledge*, *1*(1), 1–2.
Palos-Sanchez, P., Saura, J. R., & Martin-Velicia, F. (2019). A study of the effects of programmatic advertising on users' concerns about privacy overtime. *Journal of Business Research*, *96*, 61–72.
Papoutsoglou, M., Ampatzoglou, A., Mittas, N., & Angelis, L. (2019). Extracting knowledge from on-line sources for software engineering labor market: A mapping study. *IEEE Access*, *7*, 157595–157613.
Piatetsky-Shapiro, G.;1; (1989). Notes of IJCAI'89 Workshop Knowledge Discovery in Databases KDD'89. Detroit, Michigan, July.
Poddar, A., Banerjee, S., & Sridhar, K. (2019). False advertising or slander? Using location based tweets to assess online rating-reliability. *Journal of Business Research*, *99*, 390–397.
Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, California: O'Reilly Media, Inc.
Reyes-Menendez, A., Saura, J. R., & Stephen, B. T. (2020). Exploring key indicators of social identity in the #MeToo era: Using discourse analysis in UGC. *International Journal of Information Management*, *54*, 102129. http://dx.doi.org/10.1016/j.ijinfomgt.2020.102129
Ricciardi, F., Zardini, A., & Rossignoli, C. (2018). Organizational integration of the IT function: A key enabler of firm capabilities and performance. *Journal of Innovation & Knowledge*, *3*(3), 93–107. http://dx.doi.org/10.1016/j.jik.2017.02.003
Rogers, D., & Sexton, D. (2012). Marketing ROI in the era of big data. In *The 2012 BRITENYAMA Marketing in Transition Study*. pp. 1–17.
Roundy, P. T., & Asllani, A. (2018). The themes of entrepreneurship discourse: A data analytics approach. *Journal of Entrepreneurship, Management and Innovation*, *14*(3), 127–158.
Royle, J., & Laing, A. (2014). The digital marketing skills gap: Developing a digital marketer model for the communication industries. *International Journal of Information Management*, *34*(2), 65–73.
Rudder, C. (2014). *Dataclysm: Who we are (when we think no one's looking)*. Toronto, Canada: Random House Canada.
Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *4*(2), 1–40.
Saleemi, M., Anjum, M., & Rehman, M. (2017). eServices classification, trends, and analysis: A systematic mapping study. *IEEE Access*, *5*, 26104–26123.
Salminen, J., Yoganathan, V., Corporan, J., Jansen, B. J., & Jung, S. G. (2019). Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. *Journal of Business Research*, *101*, 203–217.
Sang, J., & Xu, C. (2011). Browse by chunks: Topic mining and organizing on web-scale social media. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *7*(1), 1–18.
Saura, J. R., & Bennett, D. R. (2019). A three-stage method for data text mining: Using UGC in business intelligence analysis. *Symmetry*, *11*(4), 519.
Saura, J. R., Herráez, B. R., & Reyes-Menendez, A. (2019). Comparing a traditional approach for financial brand communication analysis with a big data analytics technique. *IEEE Access*, *7*, 37100–37108.
Saura, J. R., Palos-Sánchez, P., & Cerdá Suárez, L. M. (2017). Understanding the digital marketing environment with KPIs and web analytics. *Future Internet*, *9*(4), 76.
Shareef, M. A., Kapoor, K. K., Mukerji, B., Dwivedi, R., & Dwivedi, Y. K. (2020). Group behavior in social media: Antecedents of initial trust formation. *Computers in Human Behavior*, *105*, 106225.
Song, G. Y., Cheon, Y., Lee, K., Lim, H., Chung, K. Y., & Rim, H. C. (2014). Multiple categorizations of products: Cognitive modeling of customers through social media data mining. *Personal and Ubiquitous Computing*, *18*(6), 1387–1403.
Spiekermann, S., Korunovska, J., & Langheinrich, M. (2018). Inside the organization: Why privacy and security engineering is a challenge for engineers. *Proceedings of the IEEE*, *107*(3), 600–615.
Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168.
Su, C. T., Chen, Y. H., & Sha, D. Y. (2006). Linking innovative product development with customer knowledge: A data-mining approach. *Technovation*, *26*(7), 784–795.
Su, C. T., Chen, Y. H., & Sha, D. Y. J. (2007). Managing product and customer knowledge in innovative new product development. *International Journal of Technology Management*, *39*(1–2), 105–130.
Sundsøy, P., Bjelland, J., Iqbal, A. M., & de Montjoye, Y. A. (2014). Big data-driven marketing: how machine learning outperforms marketers' gut-feeling. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 367–374). Cham: Springer.
Tiago, M. T. P. M. B., & Veríssimo, J. M. C. (2014). Digital marketing and social media: Why bother? *Business horizons*, *57*(6), 703–708.
Troisi, O., Maione, G., Grimaldi, M., & Loia, F. (2019). Growth hacking: Insights on data-driven decision-making from three firms. *Industrial Marketing Management*.
Tsapatsoulis, N., & Djouvas, C. (2019). Opinion mining from social media short texts: Does collective intelligence beat deep learning? *Frontiers in Robotics and AI*, *5*, 138.
vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *ECIS 2009 Proceedings*.
vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems*, *37*(1).
Wang, J., Zhang, W., & Yuan, S. (2017). Display advertising with real-time bidding (RTB) and behavioural targeting. *Foundations and Trends® in Information Retrieval*, *11*(4–5), 297–435.
Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), xiii–xxiii.
Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121.
Wu, J., & Holsapple, C. W. (2013). Does knowledge management matter? The empirical evidence from market-based valuation. *ACM Transactions on Management Information Systems (TMIS)*, *4*(2), 1–23.
Wymbs, C. (2016). Managing the innovation process: Infusing data analytics into the undergraduate business curriculum (lessons learned and next steps). *Journal of Information Systems Education*, *27*(1), 61.
Yeo, J., Hwang, S. W., Kim, S., Koh, E., & Lipka, N. (2018). Conversion prediction from clickstream: Modeling market prediction and customer predictability. *IEEE Transactions on Knowledge and Data Engineering*, *32*(2), 246–259.
Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, *11*(3), 1–37.
Zhu, W. Y., Peng, W. C., Chen, L. J., Zheng, K., & Zhou, X. (2016). Exploiting viral marketing for location promotion in location-based social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *11*(2), 1–28.